Vers une prise en compte de plusieurs aspects des besoins d'information dans les modèles de la recherche documentaire : Propagation de métadonnées sur le World Wide Web

Camille Prime-Claverie

▶ **To cite this version:**

Camille Prime-Claverie. Vers une prise en compte de plusieurs aspects des besoins d'information dans les modèles de la recherche documentaire : Propagation de métadonnées sur le World Wide Web. Synthèse d'image et réalité virtuelle [cs.GR]. Ecole Nationale Supérieure des Mines de Saint-Etienne; Université Jean Monnet - Saint-Etienne, 2004. Français. <NNT : 2004EMSE0020>. <tel-00839565>

HAL Id: tel-00839565

https://tel.archives-ouvertes.fr/tel-00839565

Submitted on 28 Jun 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N^o d'ordre : 342 ID

THÈSE

Présentée par Camille PRIME-CLAVERIE

pour obtenir le titre de

Docteur

DE L'UNIVERSITÉ JEAN MONNET DE SAINT-ETIENNE et DE L'ÉCOLE NATIONALE SUPÉRIEURE DES MINES DE SAINT-ETIENNE

Spécialité Informatique, Sciences de l'Information et de la Communication

Vers une prise en compte de plusieurs aspects des besoins d'information dans les modèles de la recherche documentaire : Propagation de métadonnées sur le World Wide Web

Composition du jury :

Monsieur	Yves CHIARAMELLA	Président, rapporteur	Université Joseph Fourier
Monsieur	Alain LELU	Rapporteur	Université de Franche-Comté
Monsieur	Mathias GERY	Examinateur	Université Jean Monet
Messieurs	Jean-Jacques GIRARDOT	Directeur de thèse	ENSM-SE
	Michel BEIGBEDER	Directeur de recherche	ENSM-SE
	Thierry LAFOUGE	Co-directeur de thèse	Université Lyon 1

Soutenue le 26 novembre 2004 à Saint-Etienne

Remerciements

Je tiens à remercier tout particulièrement Monsieur Thierry Lafouge et Monsieur Michel Beigbeder pour la qualité de leur encadrement. Leur complémentarité scientifique et leurs qualités humaines m'ont apporté un soutien précieux pour l'élaboration de cette thèse. Thierry, à l'écoute depuis mon stage de DEA, m'a fait découvrir avec enthousiasme le domaine de la bibliométrie. Je le remercie de m'avoir encouragée à poursuivre dans ce domaine de recherche. Auprès de Michel, j'ai toujours trouvé des conseils pertinents et je le remercie plus spécialement pour son aide en informatique.
Je les remercie tous les deux chaleureusement pour leur disponibilité, leurs lectures critiques et surtout leurs encouragements tout au long de ce travail.

Je remercie également Monsieur Jean-Jacques Girardot pour la confiance qu'il m'a accordée en acceptant de diriger mon travail, et pour m'avoir accueillie au sien du laboratoire RIM de l'Ecole nationale supérieure des mines de Saint-Étienne.

Je remercie vivement, Monsieur Yves Chiaramella et Monsieur Alain Lelu pour avoir accepté d'être rapporteurs de ma thèse. Leurs lectures attentives et leurs critiques constructives ont contribué à l'amélioration de ce mémoire.

Je remercie aussi Monsieur Mathias Géry pour l'intérêt porté à ce travail en acceptant d'être membre du jury.

Je tiens à remercier chaleureusement,

Michel et Elise Zitt pour m'avoir initiée à la scientométrie dans le cadre de mon stage de DEA et plus particulièrement à l'analyse des citations. Les discussions partagées ces dernières années m'ont beaucoup apporté tant sur le plan scientifique que personnel.

Tous les membres des équipes RIM et URSIDOC pour leur soutien et leurs conseils lors de mes présentations.

Tous les membres du centre G2I pour leur accueil, et particulièrement Madame Marie-Line Barnéoud, secrétaire du centre G2I.

Tous ceux qui m'ont accompagnée pendant ces années de thèse, et plus particulièrement Faiza, Gildas, Laurent, Hélène, Stéphane, Pascal, Franck et Annabelle.

Anne-Gaëlle, Armelle, Blandine, Emilie, Florence et Sandrine pour leur amitié.

Enfin, je souhaite remercier,
Mes parents,
Mon mari Miguel et ma petite Louise.

Table des matières

1 Introduction **1**

1.1 Le Web, un espace d'auto-publication 2

1.2 Les outils de recherche d'information sur le Web et leurs limites . 3

1.2.1 Les systèmes de recherche d'information 3

1.2.2 Les limites des moteurs de recherche 6

1.3 La représentation des documents et l'usage de métadonnées . . . 7

1.3.1 Métadonnées : définitions et origine 7

1.3.2 Usage actuel des métadonnées sur le Web 9

1.4 Prise en compte de l'hétérogénéité par les outils de recherche . . 10

1.5 Objectif et contribution de la thèse 11

1.6 Organisation de la thèse 13

2 L'analyse des liens **15**

2.1 Les réseaux sociaux 16

2.1.1 Les concepts fondamentaux en analyse des réseaux sociaux 17

2.1.2 Quelques notions de théorie des graphes 18

2.1.3 Centralité et prestige 20

2.1.4 Détection de sous-réseaux cohésifs 23

2.1.5 Pour conclure 25

2.2 La bibliométrie citationniste 25

2.2.1 L'analyse des citations : motivations et origine 26

2.2.2 Le graphe de citation et ses propriétés 28
2.2.3 Les méthodes d'analyse du graphe de citation 31
2.2.4 Les limites de l'analyse des citations 38
2.3 Le graphe du Web 40
2.3.1 Topologie et mesure du graphe 41
2.3.2 La Webométrie 46
2.3.3 Analyse du graphe et applications dans le cadre de la RI . 48

3 Notre approche de caractérisation des documents 55
3.1 Notions de documents, de sites 55
3.2 Extraction de corpus homogènes 56
3.2.1 L'hypothèse d'une auto-organisation de la Toile 57
3.2.2 Relations intéressantes dans le graphe du Web 59
3.2.3 Limites 61
3.3 Qualification des pages et des sites web 63
3.3.1 Choix des métadonnées 63
3.3.2 Proposition d'une typologie des sites et pages web 65
3.4 Conclusion 68

4 Extraction de corpus homogènes 69
4.1 Objectifs du chapitre 69
4.2 Description du protocole 70
4.3 Constitution du corpus de test 71
4.3.1 Présentation de la méthode utilisée 71
4.3.2 Etape 1 : Construction du corpus initial 74
4.3.3 Etape 2 : Recherche des prédécesseurs 75
4.3.4 Etape 3 : Réduction de la Table relationnelle *lien-entrant* 76
4.4 Qualification manuelle du corpus de test 78
4.5 Structuration du corpus par la méthode des *co-sitations* 80

4.5.1 Calcul de la matrice de *co-sitation* 81
4.5.2 Calcul de la similarité entre les URLs 84
4.5.3 Regroupement des URLs en aggrégats 88
4.6 Analyse de l'homogénité des clusters 90
4.6.1 Notions d'entropie et de redondance 91
4.6.2 Etude du pouvoir organisateur de la classification métadonnée par métadonnée 93
4.6.3 Homogénéité globale des agrégats 98
4.7 Discussion de l'expérience 100

5 Propagation de métadonnées 103

5.1 Introduction 103
5.2 Présentation des deux méthodes de propagation 104
5.3 Méthode 1 107
5.3.1 Présentation de la méthode 107
5.3.2 Algorithme de la méthode 1 108
5.3.3 Evaluation 108
5.3.4 Discussion et limites 113
5.4 Méthode 2 114
5.4.1 Présentation de la méthode 114
5.4.2 Algorithme de la méthode 115
5.4.3 Evaluation 115
5.5 Comparaison des méthodes 120
5.6 Pour conclure 120

6 Conclusion 123

6.1 Problème abordé 123
6.2 Contributions 124
6.3 Limitations 126

6.4 Perspectives : vers un passage à l'échelle 126

6.4.1 La collection 127

6.4.2 La découverte du graphe 127

6.4.3 Les caractéristiques du graphe de *sitation* 128

6.4.4 Vers la construction du graphe de *co-sitation* 130

6.4.5 Vers l'extraction de corpus homogènes 132

6.4.6 Vers l'intégration de notre approche par des outils de recherche 133

A Résultats de la qualification manuelle 135

B Listes des agrégats obtenus par la méthode du lien complet 157

Table des figures

1.1 Représentation d'un SRI traditionnel 4

2.1 Trois exemples de réseaux en fonction de leur centralité de groupe décroissante 23

2.2 Exemples d'une clique, de 2-cliques et d'un 2-clan 24

2.3 Les deux manières d'appréhender l'analyse des citations 28

2.4 le graphe de citation : graphe orienté, unidirectionnel et sans circuit 29

2.5 Distribution des citations émises (courbe réalisée à partir d'un corpus de 13.000 articles) 30

2.6 Cœur et dispersion des lois hyperboliques 31

2.7 Calcul des facteurs d'impact 32

2.8 Sous-graphe de citation pour les périodes T_1 et T_2 33

2.9 Construction d'un graphe de couplage à partir d'un graphe de citation 35

2.10 Construction d'un graphe de co-citation à partir d'un graphe de citation 37

2.11 Estimation de la taille du Web selon Lawrence and Lee Giles (1998) 42

2.12 Distribution des degrés entrant et sortant d'après Broder et al. (2000) 44

2.13 La théorie du nœud papillon selon Broder et al. (2000) 45

2.14 Exemples de graphes bipartis 52

3.1 Trois relations possibles entre les pages web 59

4.1 Etapes du processus expérimental 71

4.2 Construction du graphe de co-citation dans le cadre d'un étude bibliométrique traditionnelle 73

4.3 Etapes de la formation de notre corpus 74

4.4 Représentation de la table lien-entrant 76

4.5 Distribution des citations reçues en échelle log/log 77

4.6 Distribution des degrés entrants dans le corpus final 78

4.7 Extrait de la matrice de *co-sitation* entre les URLs 18 à 24 82

4.8 Distribution des forces de *co-sitation* 82

4.9 Distribution des degrés sortants pour les prédécesseurs 86

4.10 Extrait de la matrice de similarité entre les URLs 18 à 24 avec l'indice d'équivalence 86

4.11 Extrait de la matrice de dissimilarité entre les URLs 18 à 24 avec l'indice d_1 87

4.12 Extrait du graphe de *co-sitations* valué par l'indice de dissimilarité d_1 87

4.13 Exemple d'un dendrogramme 88

4.14 Exemple d'un effet de chaîne 89

4.15 Histogramme des valeurs du système pour le cas H_{max} 93

4.16 Distribution en rang décroissant des probabilités de formation des agrégats 96

5.1 Propagation de métadonnées selon Marchiori 104

5.2 Exemple d'un graphe valué 106

5.3 Représentation de la qualité pour les trois statégies (méthode 1) . 111

5.4 Représentation de la performance en fonction de la qualité pour les 3 stratégies (méthode 1) 112

5.5 Représentation du taux de qualification en fonction de la performance pour les 3 stratégies (méthode 1) 113

5.6 Résultats pour la stratégie du lien moyen en posant $\alpha = 5$ (méthode 1) 114

5.7 Représentation de la qualité pour les trois statégies (méthode 2) . 117

5.8 Représentation de la performance en fonction de la qualité pour les 3 stratégies (méthode 2) 118

5.9 Représentation du taux de qualification en fonction de la performance pour les 3 stratégies 119

5.10 Performance en fonction de la qualité : résultats comparatifs des deux méthodes de propagation 121

5.11 Taux de qualification en fonction de la performance : résultats comparatifs des deux méthodes de propagation 122

6.1 Distribution du nombre de pages et du nombre de netpaths par site 129

6.2 Distribution du nombre de points d'entrée par sites 129

6.3 Nombre de points d'entrée en fonction du nombre de netpaths par site 130

6.4 Distribution des fréquences de citations reçues par les points d'entrée de deux sites 131

6.5 Distribution des degrés entrant et sortant dans le corpus (pour les netpaths) 132

Liste des tableaux

4.1 Caractéristiques du fichier *corpus-initial* 74

4.2 Distribution du nombre de pages retrouvées par site 75

4.3 Caractéristiques du fichier *lien-entrant* 75

4.4 Caractéristiques du fichier *lien-entrant-ext* 78

4.5 Résultat de l'indexation 81

4.6 Exemple d'un agrégat 92

4.7 Valeurs d'entropie et de redondance pour le corpus 94

4.8 Résultats du test d'homogénéité 95

4.9 Répartition des agrégats en fonction de leur probabilité d'apparition 97

4.10 Résultats du test d'homogénéité pour les agrégats de faible probabilité d'apparition 97

4.11 Distribution de l'ordre moyen régnant dans les agrégats 99

4.12 Exemple d'un agrégat où l'ordre est total 99

4.13 Exemple d'un agrégat où l'ordre est total 100

5.1 Répartition des valeurs de métadonnées 109

6.1 Extensions les plus représentées dans la collection 127

Chapitre 1

Introduction

La fin des années quatre-vingt dix est marquée par l'apparition et l'évolution fulgurante d'un nouveau média de communication, le Web[1]. En France, 1996 représente une année essentielle dans l'histoire du Web français, celle de son ouverture au grand public. Quatre ans après, date à laquelle nous commençons ce travail de recherche, la Toile française contient plus de 85 000 sites[2], tandis que la Toile d'Araignée Mondiale est estimée à plus d'un milliard de pages. Les causes de ce succès sont nombreuses : facilité d'édition, diversité des supports (texte, images, son), couverture mondiale. Le Web se présente alors comme un véritable magma d'information regroupant des ressources très hétérogènes. La surabondance du volume d'information disponible sur la Toile entraîne inévitablement des difficultés d'accès à l'information, et des difficultés d'assimilation de l'information pour ses utilisateurs. Comment retrouver de l'information pertinente ? Comment se repérer sur la Toile ?

Ce présent travail de recherche se penche sur la recherche d'information sur le Web, et plus précisément sur les systèmes prenant en compte l'hétérogénéité de ce nouveau média. L'objectif de ce chapitre est de présenter la problématique de cette thèse. Dans un premier temps, nous analyserons les caractéristiques bien particulières de cet univers qui est considéré comme un gigantesque corpus documentaire pour certains alors qu'il n'est finalement qu'un espace d'édition non contrôlé. Nous montrerons comment cette confusion est à la base des difficultés rencontrées par les moteurs de recherche et donnerons quelques pistes pour les surmonter. Nous terminerons ce premier chapitre en présentant notre contribution ainsi que le plan de cette thèse.

[1] *World Wide Web*, en français Toile d'araignée mondiale (TAM)

[2] source AFNIC pour octobre 2000 à l'adresse `http://www.toulouse-renaissance.net/c_outils/c_statweb.htm` consulté le 19/11/03. Le Web français compte aujourd'hui 172 992 noms de domaine `http://www.afnic.fr/statistiques/afnic/afnic-repart.html` consulé le 19/11/03

1.1 Le Web, un espace d'auto-publication

Si le Web connaît aujourd'hui un tel succès auprès du grand public, et s'il a pu se développer de façon exponentielle ces dernières années, c'est avant tout grâce aux caractéristiques techniques du réseau qui le supporte. En effet, le Web est l'un des services[3] disponibles depuis 1989 sur le réseau des réseaux, appelé *Internet*. L'internet est né dans les années soixante au sein d'un organisme militaire américain, l'ARPA (Advenced Research Project Agency). Il s'est développé par la suite dans le milieu universitaire. L'origine de ce projet est la construction d'un réseau informatique capable de résister à d'éventuelles attaques soviétiques, et pouvant s'auto-configurer si l'un des maillons venait à défaillir. Emerge alors l'idée de créer un réseau *décentralisé*, *réparti* et *ouvert*. Dans le langage informatique, la décentralisation signifie que la gestion du réseau n'est pas faite par un centre de contrôle : les décisions concernant le routage des données sont prises au niveau local grâce à un algorithme distribué. Par répartition, on entend le fait que chacune des machines du réseau peut être à la fois client ou serveur. Les données sont donc stockées sur plusieurs ordinateurs. Enfin, l'ouverture permet à toute nouvelle machine de se relier au réseau quel que soit son système d'exploitation[4]. L'Internet est donc est réseau en perpétuelle expansion n'appartenant à personne en particulier puisqu'il n'a pas de structure centralisée.

A l'époque de l'*Arpanet*, le réseau ne propose qu'un service de messagerie. A partir des années soixante-dix, il devient *Internet* et regroupe les réseaux de différentes universités. Deux événements importants vont contribuer à son succès. Premièrement, l'avènement de la micro-informatique, qui va permettre au plus grand nombre de se procurer un ordinateur. Deuxièmement, la naissance du service World Wide Web développé par des chercheurs du CERN à Genève. Ce service est particulièrement attrayant car il propose un langage d'édition électronique relativement simple (le langage HTML) permettant de créer des documents hypermédias, c'est-à-dire des documents hypertextes contenant plusieurs type d'information : texte, image, sons, vidéos, etc.

L'ouverture et la distribution du Web offrent donc la possibilité à tout un chacun de mettre des documents en ligne. Très facilement, chaque individu peut devenir auteur et être lu en tout point de la planète. On assiste alors au phénomène de *l'auto-publication*. Si l'on définit le processus de publication comme *« la mise en forme d'un contenu préalablement sélectionné, en vue de sa diffusion collective »*[5], sur le Web on parle d'auto-publication car la fonction de sélection

[3] Les autres services disponibles sont le courrier électronique (e-mail), les forums de news, le transfert de fichiers (ftp), la connexion à distance (telnet), le partage de ressources documentaires (gopher).

[4] L'Internet est un réseau hétérogène supportant aussi bien, des stations Unix, des PC sous windows, des Macs, etc.

[5] Définition donné par Gabriel Gallezot sur le site de l'action spécifique pour le RTP-Documents / Département STIC du CNRS / Modèle(s) de publication sur le web. Document consulté le 3 novembre 2003 à l'adresse `http://www.unice.fr/urfist/Pubweb/texteAS.html`

n'est pas réalisée. L'édition sur le Web ne respecte en rien les contraintes et les règles de l'édition traditionnelle. Cette dernière met en relation des acteurs bien définis comme les auteurs, les éditeurs, les imprimeurs, les distributeurs, les libraires, les bibliothèques et les lecteurs dans le cas du livre. En sélectionnant les contenus, les éditeurs se positionnent en garants et responsables des informations publiées. Sur le Web la chaîne d'édition est très courte, et ne comporte que deux ou trois types d'acteurs : des auteurs, des webmestres (qui souvent sont aussi les auteurs) et des lecteurs. Si les webmestres ou les autorités hébergeant les sites sont considérés comme responsables des contenus, ce n'est que d'un point de vue juridique, pour éviter une utilisation illégale du Web. Se pose alors le problème de la qualité sur le Web, vaste programme de recherche, qui s'intéresse à la fois à la validité des informations disponibles, au contenu et à la forme des sites publiés (ergonomie, navigation, fonctionnalité)[Olsina et al., 2001]. Autre différence avec l'édition traditionnelle, l'édition sur la Toile n'impose aucun référencement systématique des documents publiés ; il n'y a rien d'analogue au dépôt légal.

Paradoxalement, le phénomène de l'auto-publication fait le succès de ce nouveau média aussi bien du côté des auteurs que des lecteurs. Les premiers en apprécient la facilité d'édition, les seconds la richesse et la diversité des contenus : diversité de langue, de culture, des sujets traités, etc. Finalement, le caractère inégal du Web et de sa qualité est admis par les lecteurs, et il devient une source d'information incontournable pour les « internautes »[6].

1.2 Les outils de recherche d'information sur le Web et leurs limites

1.2.1 Les systèmes de recherche d'information

Pour accéder facilement à l'information disponible sur la toile, le modèle classique de la recherche d'information a été transféré au Web notamment avec les moteurs de recherche et les annuaires. Un Système de Recherche d'Information (SRI) est un outil informatique qui met en relation l'expression d'un besoin d'information d'un utilisateur et un ensemble de documents considérés comme pertinents pour répondre à ce besoin. Les quatre éléments essentiels d'un SRI sont donc, le document, l'utilisateur, le système d'indexation et le système d'appariement (fig. 1.1).

1.2.1.1 Le document

Le document est l'élément central d'un système de recherche d'information. Pourtant, vouloir le définir, surtout à l'ère de l'édition numérique est une

[6] *internaute* : personne ayant accès aux différents services de l'Internet

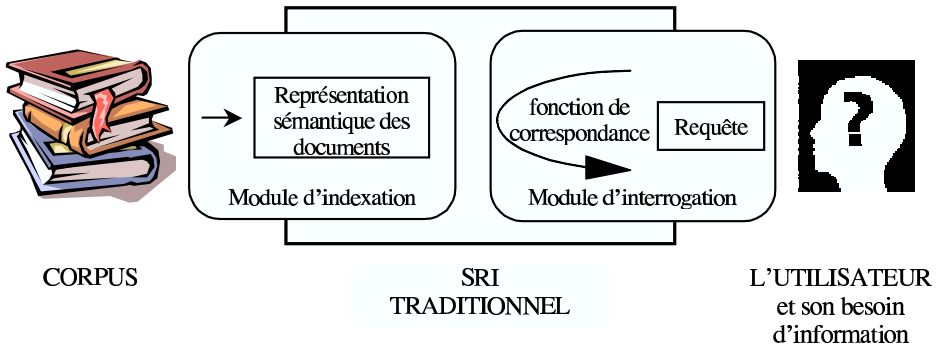

FIG. 1.1 – Représentation d'un SRI traditionnel

question délicate et peut-être vaine. Néanmoins, il est souvent considéré comme une entité matérielle véhiculant de l'information. D'un point de vue fonctionnel, il peut être recherché, retrouvé et consulté. Pour Paul Otlet, le document prend son sens lorsqu'il appartient à une collection. Il le définit comme tout objet d'une collection apportant une signification lorsqu'on l'observe. Ainsi, un document peut être un livre d'une bibliothèque, un tableau dans un musée ou encore un timbre appartenant à un collectionneur.

1.2.1.2 Le système d'indexation

Au sein d'un SRI, ce n'est pas directement le document qui est exploité mais une *représentation* de celui-ci. La représentation d'un document est une description sémantique de son contenu, généralement une liste de mots-clés appelés *descripteurs*. L'opération par laquelle on affecte des descripteurs à un document s'appelle l'*indexation* et peut être effectuée de façon humaine ou automatique :

- dans le cas de l'indexation humaine, l'indexeur doit à la fois considérer le document comme une entité en soi et envisager dans la mesure du possible, les utilisations qui pourront en être faites. Il tente de représenter au mieux le contenu du document en répondant aux questions : de quoi parle le document ? Quelle est l'information véhiculée ? Quelle est l'apport du document par rapport aux autres documents de la collection ?
- l'indexation automatique consiste à caractériser automatiquement le contenu des documents par traitements informatiques utilisant la linguistique et les statistiques. La plupart des méthodes tentent d'extraire du document les termes considérés comme le représentant au mieux, pour en faire des descripteurs. La méthode d'indexation généralement utilisée par les moteurs de recherche du Web est le *texte intégral*. Dans cette méthode tous les termes sont sélectionnés, sauf les mots-vides, termes dénués de sens : les articles, les conjonctions, les auxiliaires,

etc. Ensuite, différentes opérations linguistiques peuvent être appliquées comme la lemmatisation. Cette opération consiste à réduire un mot à sa forme canonique, c'est-à-dire supprimer toutes les variantes flexionnelles du mot liées à son usage (genre, nombre, personne). La forme canonique d'un mot correspond à son entrée dans le dictionnaire : infinitif pour les verbes, singulier pour les mots.

Qu'elle soit accomplie automatiquement ou de façon humaine, il existe trois modèles d'indexation :

- l'*indexation dite à plat* où chacun des descripteurs représente le document avec la même valeur,
- l'*indexation pondérée* où des poids sont attribués aux descripteurs en fonction de leur importance. Dans le cas du texte intégral, les poids des termes sont calculés par un algorithme qui le plus souvent prend en compte la fréquence et le pouvoir de discrimination du terme. Un terme au sein d'un document doit voir sa pondération augmentée, s'il est très fréquent dans ce document, ou s'il n'apparaît presque pas dans les autres documents. Autrement dit, un terme présent dans la plupart des documents a un faible pouvoir discriminatoire.
- l'*indexation structurée* où certaines relations entre descripteurs sont explicitées.

1.2.1.3 L'utilisateur

D'après Belkin [Belkin et al., 1982] un utilisateur va avoir recours au système de recherche d'information lorsqu'il est dans un « état anormal de connaissances » (*Anomalous States of Knowlegde*). Très souvent, cet état est déclenché par l'arrivée d'une nouvelle information qui ne peut être assimilée faute de connaissance sur le sujet. L'utilisateur entreprend alors une recherche documentaire qui commence par l'évaluation de son besoin d'information. Définir son besoin est une opération délicate, car l'utilisateur ne sait pas à l'avance ce qu'il va trouver. C'est pourquoi, une recherche documentaire est généralement un processus composé de plusieurs interrogations successives du SRI, permettant à l'utilisateur d'affiner peu à peu son besoin d'information et de l'exprimer correctement.

Comme pour le document, ce n'est pas le besoin d'information qui est directement traité mais une représentation de celui-ci. Le besoin d'information est traduit par une requête plus ou moins compliquée suivant les systèmes. Il s'agit généralement d'une suite de mots, parfois connectés par des opérateurs booléens (ET, OU, SAUF) ou par des opérateurs de proximité. Cela peut être aussi une phrase ou un petit paragraphe.

1.2.1.4 Le système d'appariement

Le système d'appariement est un intermédiaire entre l'utilisateur et les documents. Sa fonction est de sélectionner des documents pour l'utilisateur. Ce système s'appuie à la fois sur les *fichiers inverses* de descripteurs et sur la *fonction d'appariement* implémentée dans le système.

- un *fichier inverse* (appelé aussi *index*) est une table qui pour chaque descripteur donne la liste des documents qu'il décrit. Ceci permet de repérer rapidement quels sont les documents représentés par les termes de la requête. Le fichier inverse donne parfois d'autres renseignements. Dans le cas de l'indexation en texte intégral, on retrouve aussi la fréquence d'apparition du descripteur dans le document (permettant le calcul des pondérations) et sa ou ses position(s) dans celui-ci (pour l'utilisation des opérateurs de proximité).
- La *fonction d'appariement* (appelée aussi *fonction de correspondance*) a pour objectif d'attribuer un score de pertinence à chacun des documents du corpus. Suivant le modèle de recherche d'information utilisé le score peut être un booléen (pertinent ou non pertinent) ou une note variant entre 0 et 1 en fonction du degré de pertinence. Les trois modèles particulièrement connus en Recherche d'information et présentés dans l'ouvrage de référence *Information Retrieval* [van Risjbergen, 1979] sont le modèle *booléen*, le modèle *vectoriel* et le modèle *probabiliste*. Concernant les moteurs de recherche du Web, c'est le modèle vectoriel qui est le plus utilisé.

1.2.2 Les limites des moteurs de recherche

Sur le Web, l'objectif des outils de recherche généralistes (moteurs, annuaires) est double :

- parcourir l'ensemble des pages web afin d'en indexer leur contenu. Par cette tâche, les moteurs constituent leurs corpus de documents. Elle est effectuée par des robots qui fonctionnent de deux manières possibles : soit leurs parcours sont accomplis au hasard en suivant les liens hypertextes les uns après les autres (*crawler*) ; soit au contraire la liste des pages à visiter est prédéfinie : elle contient les URLs des pages à mettre à jour ou de nouvelles pages que les auteurs veulent faire connaître.

- retrouver les documents pertinents à un besoin d'information. Cet objectif est tout à fait similaire à celui des SRI présentés dans la section précédente. Dans le cas des annuaires[7] l'indexation est manuelle, alors que pour les moteurs de recherche elle se fait généralement en texte intégral.

[7] Un exemple bien connu d'annuaire est le site Yahoo (`http://fr.yahoo.com/`) qui présente les pages ou sites web par catégories thématiques.

Les outils de recherche s'appuient donc sur les techniques des SRI traditionnels ; ils en différent pourtant en un point important. Les corpus de documents au cœur des SRI traditionnels sont des *collections*. Nous définissons une collection comme un ensemble de documents sélectionnés et rassemblés par une même autorité (un documentaliste, une institution par exemple). L'intégration d'un nouveau document au sein d'une collection occasionne donc un jugement de valeur qui détermine si celui-ci remplit ou non les conditions nécessaires pour en faire partie. Les collections constituent des ensembles cohérents et homogènes où les documents partagent des propriétés communes (collections d'articles scientifiques, de brevets, etc.). D'autre part, dans les collections, les caractéristiques des documents sont généralement connues ; elles sont le résultat de l'indexation et du catalogage. Or, nous connaissons le soucis d'exhaustivité des outils de recherche généralistes que l'on rencontre sur la Toile. Le taux de pages indexées est d'ailleurs pour eux un critère de performance. Ils récoltent et indexent alors les pages web au fur et à mesure de leurs « crawls » sans aucune sélection. Leur corpus ne sont pas des collections au sens documentaire du terme.

Ces outils de recherche se heurtent à deux difficultés majeures. La première, bien connue en Recherche d'information lorsque l'on travaille sur du vocabulaire non contrôlé comme le texte intégral, concerne les problèmes issus de la langue et du langage tel que la synonymie et la polysémie entraînant des phénomènes de bruit et de silence. La seconde est directement liée au caractère hétérogène des corpus. Outre le problème de pertinence thématique, les documents rendus par les moteurs ne sont pas toujours en adéquation avec les attentes de l'utilisateur. Document trop généraliste, ou contraire d'un niveau élevé, d'un genre différent de celui attendu par l'utilisateur, etc. Il paraît donc nécessaire de ne pas se limiter à la représentation sémantique d'un document, mais de considérer aussi ses autres propriétés, comme son niveau, son origine géographique, son type (ou son genre), etc. Ce problème, précédemment soulevé par Gravano [Gravano, 2000] ne semble pas pris en compte par les outils de recherche généralistes qui ne proposent qu'un accès thématique aux documents.

1.3 La représentation des documents et l'usage de métadonnées

1.3.1 Métadonnées : définitions et origine

Dans cet univers hétérogène, l'utilisation de *métadonnées* apparaît comme la solution miracle aux difficultés de recherche d'information. Rappelons que littéralement une métadonnée est une donnée sur une donnée. Plus précisément, on peut définir les métadonnées d'une ressource comme un ensemble d'informations la décrivant et utiles pour son utilisation. A l'origine, le concept de métadonnées provient du monde des bibliothèques et de la documentation, des

archives et des musées. Au XIX^{ème} siècle avec l'apparition du catalogage moderne, il s'agit de pouvoir identifier chaque ressource par rapport aux autres de la collection. Le catalogage donne le signalement « extérieur » de chacune des ressources, c'est-à-dire les éléments la caractérisant comme son titre, ses auteurs, son éditeur commercial, sa date d'édition dans le cas du livre. Plus la collection est grande, plus le catalogage doit être précis afin d'éviter les risques de confusion entre les ressources. Les données provenant de l'indexation (section 1.2.1.1) sont aussi des éléments de métadonnées [8].

L'emploi des métadonnées s'est considérablement répandu ces dernières décennies notamment avec l'avènement de l'informatique et de l'édition numérique. Ces deux événements ont d'ailleurs contribué à faire évoluer leurs usages. Au départ conçues pour s'adresser à des êtres humains, comme les utilisateurs finaux mais aussi les différents intermédiaires, elles peuvent être aujourd'hui directement traitées par des machines. L'utilisation de standards, comme le format UNIMARC permettant l'échange de données entre bibliothèques, devient indispensable.

Dès les années 80, les chercheurs en littérature et sciences humaines prennent conscience des opportunités offertes par l'informatique pour éditer des documents électroniques [Richy, 2002] et forment un groupe de réflexion autour de cette question. Le TEI (Text Encoding Initiative) aboutit à la définition d'un standard pour les textes numérisés. Selon ce standard, un document électronique se compose de deux parties : une *en-tête* qui contient des métadonnées destinées à faciliter son utilisation, sa gestion, son indexation et son catalogage, et un *corps de document*. On parle alors de *métadonnées internes* puisque celles-ci sont directement intégrées au document, par opposition aux métadonnées stockées dans des fichiers informatiques ou papiers (*métadonnées externes*). D'autre part, émises à l'origine par des professionnels de l'information comme les documentalistes, leur provenance est aujourd'hui très diverse : métadonnées d'auteurs ou d'éditeurs et pourquoi pas directement affectées par des machines.

On distingue quatre types de métadonnées liées aux utilisations que l'on veut en faire, comme les métadonnées :

- de description, qui représentent la ressource, son apport informationnel (titre, auteur, mots-clés, etc.) ;
- administratives, liées à la gestion des ressources (propriété intellectuelle, localisation, etc.) ;
- techniques, utiles pour la consultation des ressources (données concernant la sécurité, la numérisation, etc.) ;

[8] Une différence importante entre l'indexation et le catalogage réside dans le fait, qu'il n'existe qu'une seule manière de cataloguer un document (sans erreur), alors qu'il existe de nombreuses façons d'indexer un document. En effet, les caractéristiques des documents (titre, auteurs, etc.) sont généralement non ambiguës. Par contre, l'interprétation d'un document, de l'information qu'il véhicule, dépend de l'indexeur et de ses connaissances.

– de conservation, permettant l'archivage des ressources.

1.3.2 Usage actuel des métadonnées sur le Web

Les différents langages d'édition sur le Web comme HTML et maintenant XML suivent les recommandations du TEI et prévoient l'insertion de métadonnées internes dans l'en-tête des documents. Leur utilisation est pourtant peu répandue car sans doute méconnue par la majorité des auteurs. D'autre part, ces métadonnées sont souvent mal utilisées, soit par un manque de pratique ou d'objectivité de la part des auteurs honnêtes, soit détournées de leur objectif initial pour permettre une meilleure visibilité par ceux qui les maîtrisent. C'est pourquoi finalement, la plupart des outils de recherche n'en tiennent pas compte dans leurs algorithmes de recherche. Malgré tout, des efforts de normalisation persistent. Un des standards les plus adaptés aux ressources électroniques est le projet Dublin Core [dub, 2003]. Il prévoit quinze métadonées pour décrire « bibliographiquement » les ressources de la Toile. Elles sont indépendantes du domaine d'application, et sont conçues pour décrire aussi bien des documents que des objets tels que les images, les cartes ou la musique. Ces quinze éléments de métadonnées ont trait :

– au contenu : Titre (Title), Description, Sujet (Subject), Source, Couverture spatio-temporelle (Coverage), Type, Relation avec d'autres ressources (Relation),
– à la propriété intellectuelle : Créateur (Creator), Contributeurs (Contributor), Editeur commercial (Publisher), Droits (Rights),
– à la version de la ressource : Date, Format, Identifiant (Identifier), Langage (Language).

Le standard Dublin Core décrit en dix attributs obligatoires chacun des éléments et la manière dont ils doivent être utilisés. Par exemple, un des attributs dévoile si la métadonnée est facultative ou obligatoire, un autre mentionne si elle peut avoir une ou plusieurs occurrences, etc.

Il existe actuellement une volonté de normaliser à la fois les documents numériques, les métadonnées et la sémantique des documents avec des formalismes de niveau de complexité croissante : Dublin Core (DC), Resource Description Framework (RDF), DAML+OIL (DARPA (Defense Advanced Research Projects Agency) Agent Markup Language + Ontology Inference Layer) [Richy, 2002]. L'objectif étant d'aboutir au Web sémantique, c'est-à-dire à *« un web dont le contenu peut être appréhendé et exploité par des machines. Ainsi, le web sémantique pourra fournir des services plus aboutis à ses utilisateurs (trouver l'information pertinente, sélectionner, localiser et activer le service nécessaire...) »*[9]. Ce projet conduit à la définition de standards et de langages de description de métadonnées plus ou moins complexes. Cependant les questions

[9]Source : Paris IV sorbonne, définition consulté le 14/11/03 à l'adresse `http://wiki.crao.net/index.php/WebS\%E9mantique/Quelques\%20d\%E9finitions\%20acad\%E9miques`

concernant la manière de les valuer sont trop peu évoquées.

1.4 Prise en compte de l'hétérogénéité par les outils de recherche

Deux orientations sont possibles pour surmonter les difficultés de recherche d'information liées à l'hétérogénéité du Web. La première essaye de constituer des corpus de documents homogènes (des collections) en sélectionnant un ou plusieurs types de documents bien déterminés et en n'indexant que ceux-ci. La seconde orientation, plus ambitieuse, consiste à caractériser les documents du Web pour une ou plusieurs propriétés (comme le type ou genre de document, l'origine géographique, le niveau) allant parfois jusqu'à l'affectation de métadonnées.

Concernant la première orientation, plusieurs outils de recherche spécialisés ont été créés. L'un des exemples très connu est le moteur CiteSeer[10] développé par le NEC Research Institute [Lawrence et al., 1999]. Cet outil non commercial propose d'améliorer l'accès aux ressources scientifiques disponibles sur la Toile. Son objectif initial est de localiser et d'indexer les articles scientifiques. Il offre aussi de nombreux services de navigation au sein de la littérature scientifique. La localisation des articles scientifiques ne se fait pas en parcourant le Web, mais en sollicitant de nombreux moteurs généralistes. Les interrogations utilisent les termes permettant d'accéder aux pages hébergeants des articles scientifiques souvent au format PDF ou Postscript, comme les sites de conférences. Rappelons qu'en 1999, les moteurs généralistes n'indexaient que les pages HTML, les documents écrits dans d'autres formats faisaient partie du Web invisible[11]. CiteSeer est devenu un outil de recherche particulièrement intéressant pour les chercheurs, surtout dans le domaine de l'informatique où de nombreux articles sont en ligne. Pour une meilleure visibilité sur le Web, peu à peu les auteurs soumettent directement leurs articles à ce moteur.

Concernant la seconde orientation, différents auteurs s'intéressent à la caractérisation des ressources web. Les deux propriétés particulièrement étudiées sont le genre (ou le type) des documents [Crowston and Williams, 2000], [Glover et al., 2001], [Kwasnik et al., 2001] et leur portée géographique [Ding et al., 2000]. La caractérisation des documents peut se faire soit au moment de l'indexation, soit au moment de l'interrogation.
Un exemple de caractérisation des pages dans la phase d'indexation est le processus proposé par Ding et al. [Ding et al., 2000] qui détermine la portée géographique des pages. Chaque page est alors représentée par ses termes et par

[10] `http://citeseer.ist.psu.edu/cs`
[11] CiteSeer indexe les documents PDF et Postscript en utilisant le convertiseur « ps2text »

des métadonnées géographiques. Ces méthodes implémentées dans le moteur de recherche GeoSearch[12] permettent de retrouver les articles de 300 journaux électroniques hébergés aux Etats-Unis. L'utilisateur peut préciser les Etats qui l'intéressent.
Une autre stratégie consiste à caractériser et sélectionner les documents utiles pour l'utilisateur au moment de l'interrogation. Glover et al. [Glover et al., 2001] présentent le système de recherche *Inquirus2* qui prend en compte le genre des documents. Ce métamoteur permet d'une part de reformuler automatiquement la requête de l'utilisateur en fonction de ses préférences et d'autre part d'analyser les résultats rendus pour sélectionner les documents correspondant aux attentes de l'utilisateur. Les méthodes utilisent des techniques d'apprentissage automatique (learning query modification) basées sur le contenu des pages web (texte et structure HTML) pour reconnaître certains types de documents. Le système est adapté pour retrouver les appels à communications et les « FAQ »[13].

1.5 Objectif et contribution de la thèse

Ce présent travail de recherche s'inscrit dans le cadre général de la recherche d'information sur l'Internet. Il vise la prise en compte par les outils de recherche des différents aspects des besoins d'information de l'utilisateur ; c'est-à-dire la possibilité de lui offrir le moyen d'exprimer ses besoins d'information au delà du seul critère de pertinence thématique que les anglo-saxons nomment « relevance ». En effet, les documents rendus par les moteurs même s'ils traitent du thème recherché, ne sont pas toujours en adéquation avec les attentes de l'utilisateur : document trop généraliste ou au contraire d'un niveau trop élevé, d'un genre différent de celui attendu, etc. Prenons l'exemple d'un élève et d'un chercheur espagnols recherchant tous les deux de l'information sur la physique nucléaire. Le premier s'orientera avant tout vers des mémoires ou exposés en espagnol d'un niveau vulgarisateur, alors que le second préférera des articles scientifiques probablement écrits en anglais, et pourquoi pas des appels à communication ou d'autres documents en relation avec son activité scientifique. Dans cet exemple, le type de document, son niveau et sa langue constituent des critères importants dans l'expression du besoin d'information.

L'objectif de cette thèse est donc la caractérisation des documents du Web, afin qu'ils puissent être intégrés au sein d'un système de recherche d'information évolué prenant en compte différents aspects du besoin d'information. Ce type de système comme celui décrit par le Projet ProfilDoc [Perenon, 2000] intègre à la fois :

– une modélisation de l'utilisateur et de son besoin,

[12] `http://geosearch.cs.columbia.edu/`
[13] *Frequently Asked Questions* en anglais, et « Foire aux questions » en français.

– une description approfondie des documents et de leurs propriétés matérialisées par des métadonnées.

Dans le cadre de cette thèse, nous nous plaçons du côté de la modélisation des documents, qui dans notre cas sont des pages ou sites web. Notre but est de les caractériser, ce qui en pratique peut se traduire par l'ajout de métadonnées.

Cette démarche de caractérisation des pages et sites web nous amène à nous poser trois questions :

– Quelles métadonnées utiliser pour améliorer la recherche d'information sur le Web ?
– Quelles valeurs utiliser pour chacune d'elles (vocabulaire libre, langage contrôlé) ?
– Comment les valuer ?

Répondre à la première question est particulièrement difficile car il faudrait envisager les différents besoins d'informations possibles. Cependant, les propriétés les plus importantes des documents sont sans doute celles évoquées par Gravano [Gravano, 2000] : la langue, la portée spatio-temporelle, le type/genre, la réputation et le niveau. Nous expliquons au cours du chapitre 3 pourquoi nous nous sommes plus particulièrement intéressés au genre/type des ressources web. Nous répondons en partie à la seconde question en proposant une typologie pour les pages et les sites web.

Le point-clé de notre travail est une réponse à la troisième question. L'affectation de métadonnées est une tâche délicate qui lorsqu'elle est manuelle demande du temps, une certaine maîtrise et surtout de l'objectivité. C'est pourquoi nous pensons qu'une telle opération ne peut pas être confiée directement aux auteurs. De plus, en vue d'une description uniforme et systématique des ressources, nous pensons que l'affectation des métadonnées devrait se faire du côté des systèmes de recherche d'information, de la même manière que les professionnels de la documentation effectuent le catalogage et l'indexation. Etant donné le nombre de pages disponibles sur le Web et leur volatilité, il est impossible que celles-ci soient affectées manuellement. Dans le cadre de cette thèse, nous proposons une méthode semi-automatique permettant d'affecter des métadonnées. Notre approche comporte deux étapes : l'organisation préliminaire des corpus dans l'objectif de former des sous-corpus homogènes ; l'affectation de métadonnées dans ces sous-corpus.

L'originalité de notre travail réside dans le choix du type d'information utilisé pour caractériser les documents de la Toile. La communauté du Web Mining [R and Blockeel, 2000] propose une classification présentant les trois types d'information pouvant être utilisés pour appréhender l'univers du Web. Ce sont :

– le contenu même des pages web : c'est-à-dire l'ensemble du code source de la page, le texte, les balises, les liens hypertextes, les liens vers les images ou d'autres ressources multimédias, la taille des fichiers, etc. ;

- le graphe créé par les liens hypertextes reliant les pages les unes aux autres ;
- les données provenant de l'usage comme les fichiers de log, les cookies, etc.

Remarquons que les données relatives à l'usage sont impossibles à obtenir pour l'ensemble des sites. Notre approche utilise à la fois :

- le graphe du Web, pour l'organisation des corpus. En effet, nous pensons que le graphe formé par les liens hypertextes est porteur d'information et que son analyse permet de regrouper des documents partageant des caractéristiques communes ;
- le contenu des pages, pour l'affectation des métadonnées.

Cette thèse vise donc l'étude du graphe du web en vue de l'organiser puis de caractériser les documents qu'il contient.

1.6 Organisation de la thèse

La suite de cette thèse s'organise en cinq chapitres :

- le second chapitre est un état de l'art sur le champ interdisciplinaire s'intéressant à l'analyse des liens du Web,
- le troisième présente notre approche de caractérisation des pages et argumente nos choix méthodologiques,
- les quatrième et cinquième chapitres présentent la méthode de caractérisation des pages aux travers d'expériences,
- le dernier s'intéresse à la mise en pratique d'une telle méthode sur un corpus de taille importante, en évoque les limites et ouvre des perspectives.

Chapitre 2

L'analyse des liens

Le World Wide Web est à ce jour la plus grande application de système hypertexte. La notion d'hypertexte est relativement ancienne. Elle repose sur l'idée que la strucure des documents peut être non-linéaire, et que leur lecture peut se faire suivant différents parcours. Vannevar Bush apparaît dès 1945, comme un des pères fondateurs des systèmes hypertextes. Une de ses préoccupations était déjà à l'époque celle de l'explosion documentaire. ll voulait éviter que des découvertes scientifiques soient inexploitées car perdues dans la masse des publications, comme l'ont été pendant toute une génération les travaux en génétique de Mendell. Selon Bush, un document est utile à la science s'il est complété, enrichi et augmenté, et surtout, s'il peut être facilement localisé et consultable [Teasdale, 1995]. Dans son article « As we may think » [Bush, 1945], il décrit le système imaginaire Memex (Memory extented). Il s'agit d'un système contenant une immense collection de documents sur microfilms et dont la particularité est de pouvoir associer les documents entre eux. Les associations entre les documents matérialisent différents cheminements possibles de la pensée. Compte tenu des moyens techniques de l'époque, le système Memex n'est resté qu'un modèle théorique de navigation entre les documents.

Vingt ans plus tard, alors que le professeur américain Douglas Engelbart propose une mise en oeuvre des idées de Bush notamment avec le système NLS (oN Line System), Theodor Nelson, créateur du projet Xanadu[1] invente les termes *hypertexte* et *hypermédia*. Depuis 1987, le grand public se familiarise peu à peu avec les systèmes hypertextes, notamment avec l'apparition du premier logiciel « grand public » d'édition d'hypertextes *HyperCard*, puis par la découverte du World Wide Web à la fin des années 1990.

[1]Xanadu est un projet de publication en réseau, sorte de dépôt des œuvres permettant leur consultation et la gestion du droit d'auteur et des copyright. Le système permet de gérer dans un même et unique environnement tous les documents, mais surtout, permet au lecteur de modifier et d'adapter les liens entre les documents en fonction de ses connaissances.

Un système hypertexte contient donc des nœuds d'informations reliés entre eux par des liens. Les liens permettent d'une part, d'organiser un document de manière non séquentielle, et d'autre part, de le relier à d'autres documents comportant de l'information similaire ou complémentaire. Les liens sont donc porteur de sens. Rappelons que l'objectif initial de Bush et Engelbart était de plaquer sur un ensemble d'information (des documents), un réseau de connaissances, pour que cet ensemble soit compréhensible par l'homme et la machine [Balpe et al., 1996]. Sur le Web, l'objectif fixé par Bush et Engelbart n'est certainement pas atteint. En effet, les raisons de lier deux pages web sont nombreuses et ne se limitent pas au champ sémantique. C'est d'ailleurs pourquoi nous pensons que l'analyse des liens de la Toile est d'autant plus intéressante, puisque ces liens hypertextes matérialisent des relations très diverses comme des relations sociales, des relations thématiques, etc.

Depuis 1996, les travaux de recherche autour de l'analyse des liens du Web se multiplient. L'objectif principal est de mieux comprendre la structure du Web afin d'en améliorer l'accès à son contenu. Les techniques utilisées sont celles de la théorie des graphes mais se rapprochent aussi des travaux antérieurs issus de l'analyse des réseaux sociaux et de la bibliométrie.

L'objectif de ce chapitre est de présenter le champ de recherche interdisciplinaire s'intéressant à l'analyse du graphe du Web et de montrer comment les apports de plusieurs communautés s'entremêlent afin d'appréhender cet univers. Ce chapitre comporte trois sections.

- Les deux premières présentent les travaux plus ou moins anciens de l'analyse des réseaux sociaux et de la bibliométrie citationniste.
- Dans la troisième section, nous nous intéressons au graphe du Web, à sa structure et aux analogies possibles avec l'analyse des réseaux sociaux et la bibliométrie. Nous présenterons aussi les applications de l'analyse des liens dans le cadre de la recherche d'information sur le Web.

2.1 Les réseaux sociaux

L'analyse des réseaux sociaux résulte d'un effort interdisciplinaire entre les sciences sociales, les mathématiques formelles, les statistiques et l'informatique. Elle s'intéresse aux individus, aux relations qu'ils entretiennent, mais surtout aux implications de ces relations. Elle constitue un champ de recherche très intéressant pour les chercheurs en sciences sociales et comportementales qui considèrent que les acteurs et leurs actions sont interdépendants. Bien que le terme *social network* n'apparaisse qu'en 1954 [Barnes, 1954], l'idée selon laquelle les relations sociales influent de manière positive ou négative sur les individus émerge depuis le début du XX$^{\text{ème}}$ siècle en sociologie et psychologie. La sociométrie, précurseur de l'analyse des réseaux est introduite par Moreno dans

les années 30. Moreno est le premier à mettre au point une véritable méthode d'analyse des réseaux basée sur le *test sociométrique* et le *sociogramme* :

« Le test sociométrique est un instrument qui sert à mesurer l'importance de l'organisation qui apparaît dans les groupes sociaux. Il consiste expressément à demander au sujet de choisir, dans le groupe auquel il appartient ou pourrait appartenir, les individus qu'il voudrait avoir pour compagnons. On lui demande d'exprimer ses choix sans réticence, que les individus choisis fassent ou non partie du groupe actuel. Le test sociométrique est un instrument qui étudie les structures sociales à la lumière des attractions et des répulsions qui se sont manifestées au sein du groupe. » [Moreno, 1934]

Les résultats du test sociométrique sont représentés en deux dimensions sur un graphique appelé sociogramme. Sur ce graphique les individus sont matérialisés par des points et des traits montrent les relations entre eux. Cette représentation est effectuée manuellement. Plus le nombre d'individus est important, plus grande est la difficulté de les positionner pour que leurs relations apparaissent clairement. C'est grâce, conjointement aux résultats de la théorie des graphes et aux progrès informatiques que l'analyse des réseaux sociaux apparaît ces trente dernières années comme un domaine de recherche à part entière, peu présent il est vrai en France. Elle propose des moyens pour définir les concepts sociaux importants et des indicateurs mathématiques mesurant les propriétés sociales structurelles. Les méthodes s'appliquent aux différentes questions intéressant les chercheurs en sciences sociales. Quelques exemples présentés dans l'ouvrage de référence *Social network analysis* [Wasserman and Faust, 1994] sont : le monde politique et le système économique, les consensus et l'influence sociale, les communautés, la sociologie de la science. L'étude des réseaux est aussi utile dans d'autres domaines, tel que la médecine pour les travaux concernant l'épidémiologie, ou même, pour le renseignement et l'espionnage.

2.1.1 Les concepts fondamentaux en analyse des réseaux sociaux

L'analyse des réseaux sociaux concerne donc l'étude d'acteurs et de leurs relations. D'un point de vue formel, le réseau est représenté sous la forme d'un graphe $\mathcal{G}$ composé de deux ensembles d'information :

- un ensemble de nœuds ou sommets, noté $\mathcal{N} = \{n_1, n_2, ..., n_N\}$, représentant les N acteurs ;
- et un ensemble d'arêtes entre deux nœuds, noté $\mathcal{A} = \{a_1, a_2, ..., a_A\}$, représentant les A liens relationnels.

1. Les acteurs
   Nœuds du réseau, les acteurs peuvent être des personnes individuelles, un

groupe, ou un ensemble d'unités sociales. L'emploi du terme « acteurs » n'implique pas nécessairement que ceux-ci soient capable d'agir. Dans la majorité des applications, les acteurs sont du même type (individus travaillant dans la même société par exemple) et dans ce cas le réseau est qualifié « d'unimodal ». Cependant, il arrive parfois que certaines études s'intéressent à des acteurs de type différents.

2. Les liens relationnels
   Les liens relationnels sont des canaux permettant la circulation de ressources matérielles ou non. Ils peuvent indiquer des relations et des échanges très variés comme par exemple les sentiments entre les individus (l'amitié, le respect), les échanges matériels (le commerce), les mouvements entre les lieux géographiques (les migrations), etc. Ils sont parfois qualifiés par leur intensité (le nombre de produits vendus par exemple). Ces liens peuvent être :
   - *non orientés* indiquant une relation réciproque comme les relations d'amitié par exemple ;
   - *orientés* indiquant le sens de circulation de la ressource. Dans ce cas les arrêtes sont appelées *arcs*.

2.1.2 Quelques notions de théorie des graphes

L'analyse des réseaux sociaux ainsi que les différents champs de recherche présentés dans les sections suivantes reposent en partie sur la théorie des graphes. Il nous semble donc nécessaire de définir les principaux concepts de cette théorie ainsi que de proposer des notations qui seront utilisées par la suite.

Terme	Notation	Définition
Graphe (non orienté)	$\mathcal{G} = (\mathcal{N}, \mathcal{A})$	Ensemble de points appelés nœuds ou sommets dont certaines paires sont reliées par des liens nommés arêtes.
Ensemble des nœuds	$\mathcal{N}(\mathcal{G}) = \{n_1, n_2, ..., n_N\}$	Ensemble des N nœuds (sommets) du graphe $\mathcal{G}$. Le nombre de nœuds est appelé *ordre du graphe*.
Ensemble des arêtes	$\mathcal{A}(\mathcal{G}) = \{a_1, a_2, ..., a_A\}$	Ensemble des A arêtes du graphe $\mathcal{G}$
Degré d'un nœud	$d(n_i)$	Nombre de sommets adjacents à un nœud donné. n_i et n_j sont adjacents s'il existe une arête $a_k = \{n_i, n_j\}$ entre ces deux nœuds.
Chaîne	C	Suite de nœuds reliés par des arêtes
Cycle		Chaîne dont les extrémités coïncident.

Terme	Notation	Définition
Graphe orienté	$\mathcal{G} = (\mathcal{N}, \mathcal{A})$	Graphe dont les liens sont orientés et nommés *arcs*. Un arc a_k de $\mathcal{A}$ est alors défini par un couple de sommets, $a_k = (n_i, n_j)$ signifie que l'arc a_k va de n_i à n_j. n_i est appelé extrémité initiale et n_j l'extrémité finale de a_k.
Degré entrant	$d_e(n_i)$	Nombre d'arcs ayant pour extrémité finale le nœud n_i
Degré sortant	$d_s(n_i)$	Nombre d'arcs ayant pour extrémité initiale le nœud n_i
Chemin	C	Dans un graphe orienté, suite de sommets reliés par des arcs.
Circuit		Dans un graphe orienté, chemin fermé simple.
Graphe valué	$\mathcal{G} = (\mathcal{N}, \mathcal{A})$	Graphe orienté (ou non) où des nombres réels, appelés aussi poids, sont associés aux arcs (ou aux arêtes).
Distance géodésique	$d(n_i, n_j)$	Longueur de la plus courte chaîne ou du plus court chemin entre les nœuds n_i et n_j. Dans un graphe valué, somme des valuations des arêtes (arcs) de cette chaîne (chemin).
Matrice d'adjacence	$M(\mathcal{G})$	Représentation matricielle du graphe. La matrice d'adjacence est carrée. Elle est binaire pour le graphe non valué : la valeur 1 en position $(i; j)$ indique l'existence d'une arête (ou d'un arc) du sommet i au sommet j, la valeur 0 indique le contraire. Dans la cas du graphe valué, x en position $(i; j)$ indique la valuation de l'arête (ou de l'arc) du sommet i au sommet j. ∞ (ou l'ordre du graphe N) en position $(i; j)$ signifie qu'il n'existe pas d'arête (ou d'arc) du sommet i au sommet j.

2.1.3 Centralité et prestige

Une des finalités de l'analyse des réseaux est de repérer quels sont les acteurs les plus importants. Cette notion d'importance correspond à la visibilité, la proéminence des acteurs. On considère qu'un acteur est proéminent, si les liens qu'il entretient avec d'autres le rendent particulièrement visible, c'est-à-dire s'il est largement impliqué dans le réseau. La notion de proéminence tient compte non seulement des liens directs (liens adjacents) mais aussi des liens indirects impliquant des intermédiaires. On distingue deux notions pour mesurer la proéminence : la *centralité* et le *prestige*.

2.1.3.1 La centralité

La centralité tend à définir le ou les acteurs les plus au centre du réseau. Il existe plusieurs indices pour mesurer la centralité des acteurs, introduits et discutés par les différents auteurs du domaine. Pour illustrer cette notion, nous présenterons trois indices de centralité pour le cas des réseaux non orientés : la *centralité de degré*, la *centralité de proximité*, la *centralité d'intermédiarité*.

- Centralité de degré (notée C_D)
  C'est sans doute la définition la plus simple de la centralité. La centralité de degré détermine les acteurs les plus centraux comme les acteurs les plus actifs, c'est-à-dire comme ceux qui sont le plus reliés à d'autres dans le graphe. Mathématiquement, il s'agit des acteurs dont le degré est élevé :
  $$C_D(n_i) = d(n_i)\,. \tag{2.1}$$
  Cette mesure est dépendante du nombre de sommets du réseau et ne permet pas de comparer la centralité d'acteurs de réseaux différents. Comme la valeur maximale de centralité que peut prendre un acteur est le nombre maximum de relations qu'il peut entrenir avec d'autres, c'est-à-dire $N-1$, il convient de normaliser cette mesure de centralité de la manière suivante :
  $$C'_D(n_i) = \frac{d(n_i)}{N-1}\,. \tag{2.2}$$

- Centralité de proximité (notée C_P)
  La seconde approche de la centralité est basée sur la notion de distance dans les graphes. Un acteur est d'autant plus central s'il est proche de tout les autres, c'est-à-dire s'il peut rapidement interagir avec eux. Introduite par Sabibussi [Sabidussi, 1966], la mesure la plus simple consiste pour un sommet donné, à calculer sa proximité aux autres sommets. La proximité se définit comme la somme des distances d'un sommet à tous

les autres. Le sommet est considéré comme central si cette somme est faible. La centralité de proximité est donc l'inverse de la proximité :

$$C_P(n_i) = \frac{1}{\sum_{j=1}^{N} d(n_i, n_j)} . \tag{2.3}$$

L'indice C_P est maximum lorsque la somme $\sum_{j=1}^{N} d(n_i, n_j)$ atteint sa valeur minimum $N-1$, c'est-à-dire lorsque le sommet n_i est adjacent à tous les autres. Il est minimum lorsque n_i est très éloigné des autres ; dans ce cas l'indice tend vers 0. La forme normalisée de cet indice est donc :

$$C'_P(n_i) = \frac{N-1}{\sum_{j=1}^{N} d(n_i, n_j)} . \tag{2.4}$$

– Centralité d'intermédiarité (notée C_I)
Cet indice de centralité repose sur la notion d'intermédiaire. Un sommet est un bon intermédiaire s'il se situe sur de nombreux chemins géodésiques dans le graphe. Deux sommets peuvent avoir plusieurs chemins géodésiques, c'est-à-dire plusieurs « plus courts chemins » pour les relier. On note g_{jk} le nombre de chemins géodésiques entre les nœuds n_j et n_k. $g_{jk}(n_i)$ est le nombre de chemins géodésiques contenant le sommet n_i. La centralité au sens de l'intermédiarité pour le nœud n_i se calcule de la manière suivante :

$$C_I(n_i) = \sum_{j<k} \frac{g_{jk}(n_i)}{g_{jk}} \qquad avec \quad j \neq i \quad \text{et} \quad k \neq i \tag{2.5}$$

Son maximum est atteint lorsque n_i appartient à tous les géodésiques ; $C_I(n_i)$ prend alors la valeur $(N-1)(N-2)/2$. Cette valeur est le nombre de couples de sommets issus de l'ensemble $\mathcal{N}$ privé d'un sommet. Il est minimum et nul lorsque n_i n'appartient à aucun chemin. Il se normalise de la manière suivante :

$$C'_I(n_i) = \frac{2 \times C_I(n_i)}{(N-1)(N-2)} . \tag{2.6}$$

2.1.3.2 Le Prestige

Dans le cas de réseaux orientés, on peut faire la distinction entre donner et recevoir, choisir ou être choisi. Dans de tels réseaux, la centralité s'intéresse aux choix que font les acteurs (liens sortants), tandis que le *prestige* examine les choix reçus (liens entrants). La notion de prestige est analogue à la « popularité », elle est d'autant plus forte que l'acteur reçoit beaucoup.

Comme pour la centralité, il existe plusieurs indices pour mesurer le prestige. D'ailleurs certains indices de centralité sont adaptés pour mesurer le prestige en ne tenant compte que des liens et des chemins entrant sur les sommets. Par exemple, le prestige de degré se calcule en comptant le nombre de liens entrant et s'écrit comme suit :

$$P_D(n_i) = d_e(n_i)\,. \tag{2.7}$$

La forme normalisée de cet indicateur est :

$$P'_D(n_i) = \frac{d_e(n_i)}{N-1}\,. \tag{2.8}$$

2.1.3.3 Centralité et prestige de groupe

Les mesures de centralité et de prestige permettent d'identifier les acteurs proéminents au sein d'un réseau. Il convient de préciser que la présence d'acteurs centraux ou prestigieux est liée à la structure même du réseau. Certaines formes de réseaux n'impliquent pas d'acteurs proéminents, le réseau circulaire (figure 2.1) en est un bon exemple. D'autres structures, comme le réseau en étoile, créent un acteur central « parfait » en le positionnant au centre et le reliant à tous les autres. Les notions de centralité et de prestige ont donc été étendues au groupe[2] pour mesurer sa capacité à contenir des acteurs proéminents. Freeman [Freeman, 1979] propose un indice mathématique pour mesurer la centralité d'un groupe composé de g acteurs. La *centralité de groupe* tient compte des valeurs de centralité obtenues par chaque sommet du groupe et ce quel que soit la centralité choisie (centralité de degré, de proximité, d'intermédiarité).
Notons $C_a(n_i)$ la centralité d'un acteur pour la centralité a choisie. $C_a(n^*)$ est la valeur de centralité maximale prise dans le groupe, on a donc : $C_a(n^*) = max_{n_i} C_a(n_i)$. Freeman invite à calculer la somme des différences entre les valeurs de centralité prises par les acteurs du groupe avec la valeur maximale. La forme générale de la centralité de groupe est la suivante :

$$C_a(groupe) = \sum_{i=1}^{g} [C_a(n^*) - C_a(n_i)]\,. \tag{2.9}$$

Plus cette différence est importante, plus le groupe est central. Cette différence est maximun pour le réseau en étoile et minimum pour les réseaux réguliers[3], ce qui est le cas du réseau circulaire (fig. 2.1).

[2] Un groupe est un ensemble d'acteurs et de leurs relations dans le réseau étudié (sous-réseau).

[3] Un graphe d-régulier est un graphe dont tous les sommets sont de degré d.

La centralité de groupe peut être obtenue plus simplement en calculant la variance des valeurs de centralité normalisées des sommets du groupe [Boutin, 1999]. Ainsi, la valeur de centralité de goupe est indépendante du nombre de sommets ce qui permet d'effectuer des comparaisons entre les groupes.

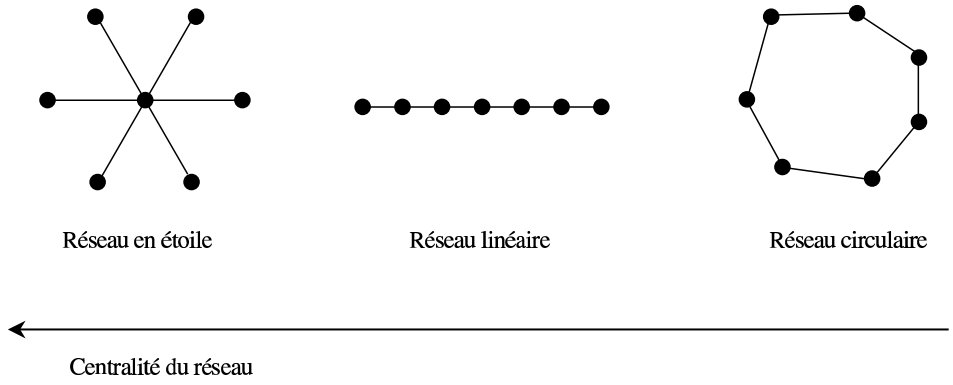

FIG. 2.1 – Trois exemples de réseaux en fonction de leur centralité de groupe décroissante

2.1.4 Détection de sous-réseaux cohésifs

Un autre axe de recherche important de l'analyse des réseaux s'intéresse à la détection de sous-réseaux fortement interconnectés. En effet, différents auteurs pensent que les relations entre les individus surtout lorsqu'elles sont positives, influent sur leurs comportements et amènent une « standardisation » du groupe. Suivant cette idée, on attend plus d'homogénéité entre deux personnes qui sont fréquemment en face-à-face, ou connectées par des intermédiaires, qu'entre deux personnes qui ne sont que très rarement (voire jamais) en relation. La cohésion de groupe est un critère pour explorer les théories sociologiques basées sur l'influence et les consensus. Les propriétés des groupes cohésifs sont les suivantes :

- la présence de relations mutuelles dans le groupe,
- la proximité des acteurs du groupe,
- l'importance (en nombre) des liaisons au sein du groupe,
- la supériorité du nombre de liaisons entre les acteurs membres du groupe par rapport aux non-membres.

Il existe plusieurs méthodes permettant de découvrir l'existence de groupes cohésifs. Toutes se basent sur les propriétés relationnelles au sein des groupes (l'adjacence, la distance géodésique ou le nombre de relations entre les membres). Différents choix peuvent être faits et les méthodes habituellement utilisées tentent de repérer des sous-réseaux répondant au moins à l'une des quatre propriétés

évoquées plus haut.

La méthode la plus ancienne est basée sur la mutualité complète et recherche la présence de *cliques* dans le graphe. Les cliques sont définies comme des ensembles de sommets tous reliés les uns aux autres. Formellement, les cliques sont des sous-graphes complets maximaux composés d'au moins trois sommets. Autrement dit, les couples de sommets mutuellement en relation ne sont pas considérés comme des cliques. Précisons que les cliques sont des définitions très strictes de groupes cohésifs. En effet, il suffit que seuls deux sommets ne soient pas en relation pour que le groupe ne soit pas considéré comme cohésif. C'est pourquoi cette notion de groupe cohésif a été étendue, prenant en compte d'autres critères comme la proximité des sommets et le diamètre du groupe.

Ainsi plusieurs types de groupes cohésifs ont été définis. Les n-cliques par exemple sont des ensembles de sommets tous distants par la distance géodésique d'au maximum n ; les n-clans sont des ensembles de sommets tous distants d'au maximum n en ne tenant compte que des chemins possibles dans le sous-graphe (fig. 2.2).

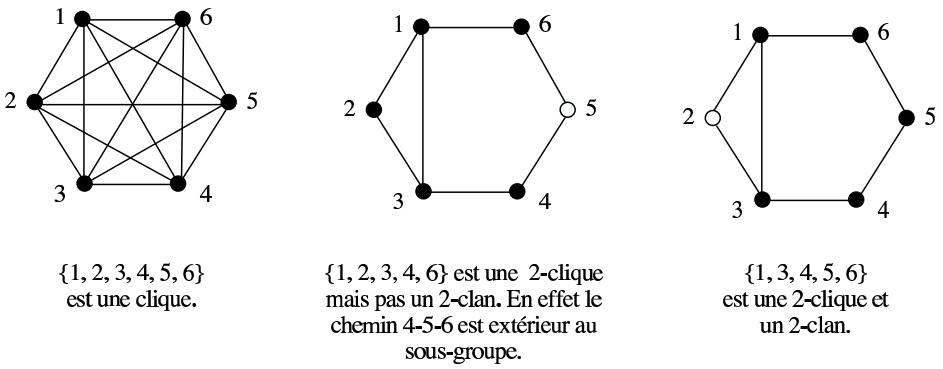

FIG. 2.2 – Exemples d'une clique, de 2-cliques et d'un 2-clan

Les k-noyaux, autre concept de groupe cohésif, se basent sur les degrés des sommets. Dans un k-noyau, les sommets sont tous adjacents au moins à k sommets du groupe. Enfin, la dernière forme de groupe cohésifs, les LSSets, tient compte de la supériorité des liaisons internes au groupe par rapport aux liaisons externes.

Nous avons présenté brièvement les différentes méthodes permettant de découvrir des sous-réseaux cohésifs basées sur des propriétés physiques du graphe. Une autre alternative consiste à utiliser des méthodes statistiques pour décou-

per les graphes en sous-réseaux cohésifs. Ces méthodes sont les MDS (Multi-Dimentional Scaling), les analyses factorielles ou les classifications. Nous montrerons dans les sections suivantes comment fonctionnent ces méthodes, plus particulièrement au chapitre 4 où nous présenterons une expérience de « découpage » d'un graphe utilisant les classifications hiérarchiques ascendantes.

2.1.5 Pour conclure

En conclusion nous pouvons dire que les méthodes développées par le champ disciplinaire de *l'analyse des réseaux sociaux* permettent d'appréhender les graphes d'acteurs de deux manières différentes. Premièrement, d'un point de vue individuel : les sommets sont analysés séparément dans l'objectif de mesurer leur implication dans le graphe (centralité, prestige). Deuxièmement, de manière collective pour tenter d'identifier des groupes d'acteurs homogènes partageant des propriétés communes. Nous retrouverons ces deux approches pour chacun des graphes que nous étudierons, que ce soit le graphe de citations en bibliométrie (section 2.2) ou le graphe du Web (section 2.3).

2.2 La bibliométrie citationniste

Le *Dictionnaire encyclopédique des sciences de l'information et de la communication* définit la bibliométrie comme « l'exploitation statistique des publications » (scientifiques). Introduit en 1969 par Pritchard [Pritchard, 1969] le concept de bibliométrie s'intègre dans la notion plus large de *scientométrie*. La scientométrie vise l'étude de l'activité scientifique avec deux motivations principales [Bassecoulard-Zitt and Zitt, 1998] :

- comprendre la science comme une activité humaine. La science devient alors objet d'étude pour les sociologues qui tentent de découvrir quelles règles régissent l'activité scientifique. Dans son ouvrage *Little Science, Big Science* [de Solla Price, 1963b], Derek de Solla Price propose de parvenir à « une compréhension globale de la croissance de la science et de son comportement d'ensemble » en traitant la science « comme une entité mesurable ». Ses recherches dans ce domaine l'amènent à proposer plusieurs lois reconnues actuellement en Sciences de l'information, comme *la loi des avantages cumulés*, « le succès engendre le succès » ou le phénomène *des collèges invisibles* induit par les règles sociologiques de collaborations entre chercheurs.

- évaluer l'activité scientifique et prévoir son évolution. Cette orientation est particulièrement intéressante pour les décideurs politiques d'une part, qui cherchent à obtenir des indicateurs de production scientifique

permettant d'allouer ainsi des crédits de recherche et pour les industriels d'autre part, qui surveillent la concurrence et orientent leurs activités de recherche.

Dans ce contexte la bibliométrie apparaît comme l'ensemble des méthodes quantitatives permettant de formaliser les règles de l'activité scientifique. Les études bibliométriques travaillent sur des corpus volumineux de publications scientifiques, généralement des articles primaires ou des brevets. D'un point de vue pratique, ceci est rendu possible grâce à l'existence et l'utilisation des banques de données bibliographiques. Les notices bibliographiques sont des ensembles structurés d'information, séparés en champs, dont certains, que nous appellerons *marqueurs*, sont particulièrement riches d'information pour contribuer à l'analyse de l'univers scientifique. Les champs *mots-clés* et *titre* en sont de bons exemples. Ils figurent d'ailleurs parmi les champs les plus souvent utilisés dans les études bibliométriques. L'ISI[4] (Institute for Scientific Information), créé par Eugène Garfield en 1963, propose 3 bases bibliographiques d'un genre particulier. Pour chaque notice, on retrouve à la suite des champs traditionnels (titre, auteur, journal, etc.), les références bibliographiques mentionnées dans l'article ainsi que les citations qu'il a reçues. Les trois bases d'index gérées par l'ISI, le *Science Citation Index*, le *Social Sciences Citation Index*, et le *Art and Humanities Citation Index*, couvrent les revues scientifiques considérées par l'éditeur comme les plus importantes de chaque domaine, dont la plupart sont en langue anglaise.

Deux approches sont donc possibles en bibliométrie : une approche lexicale et une approche citationniste. Dans l'approche lexicale, les marqueurs de sens utilisés sont des mots : en général, les mots-clés, les mots du titre ou du résumé. Les méthodes sont tributaires de la langue utilisée et se heurtent souvent aux problèmes issus de la langue et du langage (polysémie, synonymie). Dans l'approche citationniste, les marqueurs de sens sont les citations faites par les auteurs, c'est-à-dire les références bibliographiques des articles scientifiques. C'est cette dernière approche qui nous intéresse tout particulièrement, puisque les citations faites par les auteurs forment des liens entre les documents, créant un réseau de publications scientifiques semblable aux hypertextes ou aux réseaux sociaux.

2.2.1 L'analyse des citations : motivations et origine

Il est d'usage depuis le XIX$^{\text{ème}}$ siècle que le chercheur mentionne à la suite de son article l'ensemble des travaux qui l'ont aidé dans le cadre de sa recherche. Ces citations permettent d'une part, aux lecteurs de consulter les travaux qui ont inspiré l'auteur ; d'autre part, c'est aussi une façon pour lui de rendre hommage à ses prédécesseurs. Une contribution scientifique est donc un document structuré qui se compose en général de trois éléments :

[4] `http://www.isinet.com`

- un corps d'article qui relate son apport scientifique ou technologique,
- une série d'annotations ou de citations qui permet de situer le document par rapport à l'accumulation passée du savoir,
- et enfin une liste de références bibliographiques, qui reprend en général les documents mentionnés par les annotations ainsi que d'autres documents potentiellement intéressants pour le lecteur.

L'article scientifique occupe une place essentielle en bibliométrie, car il est considéré comme un indicateur de production de la recherche scientifique. Selon le *réductionnisme bibliométrique*, « point de vue par effet duquel l'article scientifique devient un outil de définition de la science et l'on fait de la publication écrite un indicateur privilégié de l'activité scientifique, [...] le produit final de la recherche scientifique est la publication d'un texte écrit » [Polanco, 1995]. Par ce point de vue, sont considérés comme scientifiques ceux qui publient. On observe d'ailleurs une certaine dérive du côté des chercheurs qui n'utilisent plus l'article scientifique dans sa fonction première, celle de communiquer leurs savoirs, mais pour se faire reconnaître et cautionner la propriété intellectuelle de leurs travaux.

Si la quantité d'articles produits a été l'un des premiers indicateurs utilisés en bibliométrie pour mesurer et représenter l'activité scientifique, Price pense que « le degré d'utilisation semble être un meilleur test de qualité » [de Solla Price, 1963a]. Le degré d'utilisation se traduit non seulement par la consultation et la circulation des articles [Lafouge, 1991] mais aussi par la fréquence de citations qu'ils reçoivent. C'est d'ailleurs, l'hypothèse de base de l'analyse des citations de Price proposée par dans *Little Science, Big Science*. Selon lui, la fréquence des citations et des références mesure « l'utilité des différents articles ».

Formellement, on entend par citation, la relation entre un document citant et un document cité dans la liste des références bibliographiques. Pour être plus précis, il faudrait faire la distinction entre références et citations. « Si l'article A a une note bibliographique utilisant et décrivant l'article B, alors A contient une référence à B, et B reçoit une citation de A » [de Solla Price, 1970]. L'analyse des citations peut être envisagée de deux manières (fig. 2.3) : cela peut être l'étude des références $A_1, ..., A_n$ citées par un article B, ou l'étude des articles $B_1, ..., B_n$ ayant des références à un article antérieur A.

Dans la situation 1, l'analyse des références citées suppose que l'ensemble des références bibliographiques d'un article B représente une source d'information exploitable, car elle permet de comprendre comment se situe le chercheur par rapport à l'accumulation passée du savoir.

Dans la situation 2, l'analyse des articles B se référant à l'article A, essaie de mesurer l'impact de l'article A, ou son utilité au sein de la communauté

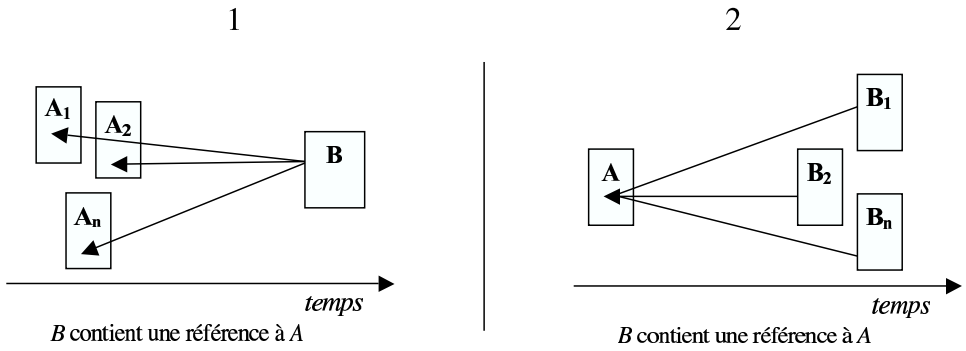

FIG. 2.3 – Les deux manières d'appréhender l'analyse des citations

scientifique [de Solla Price, 1963b].

La structure des publications scientifiques est semblable à un réseau, dans lequel les nouveaux documents sont reliés aux anciens par l'intermédiaire de la citation. Ce réseau peut être représenté par une matrice booléenne ou par un graphe orienté associé. L'article *Network of Scientific Papers* [de Solla Price, 1965] marque le début de l'analyse des citations selon Price. Celui-ci présente les citations comme un *indicateur relationnel* qui permet « de reconstruire empiriquement la dimension socio-cognitive sous-jacente d'un champ scientifique, mais aussi de la représenter visuellement sous la forme d'une carte » [Polanco, 1995]. Approuvant les idées de Price, Garfield crée à Philadelphie (USA) au début des années soixante l'ISI (Institute for Scientific Information) et sa première base d'index le Science Citation Index (SCI) en 1963. A titre indicatif, la première base contient 1,4 millions de citations issues de 613 journaux de l'année 1961. Le travail de Garfield permet donc d'explorer par l'expérience les idées théoriques de Price.

2.2.2 Le graphe de citation et ses propriétés

Le réseau des publications scientifiques peut être représenté sous la forme d'un graphe orienté $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, où les nœuds sont les publications et les arcs les relations entre elles obtenues par les citations.

1. Les publications scientifiques
   Elles sont les composantes élémentaires du modèle scientifique et appelées *items*. Elles sont datées et appartiennent à différentes *unités scientifiques* comme les auteurs, les revues, les institutions, les pays, etc.
2. Les citations
   Par l'intermédiaire des références bibliographiques elles relient les différents items. De manière indirecte, elles relient aussi les différentes unités scientifiques.

2.2.2.1 La dimension temporelle du graphe de citation

Une des particularités du graphe de citation est son aspect dynamique. Il n'est pas stable mais est en évolution perpétuelle. A chaque fois qu'un nouvel article est publié il est « raccordé » au graphe par l'intermédiaire de ses références bibliographiques. L'apparition d'un nouvel article ne modifie pas les relations existantes, mais en ajoute de nouvelles. Une des conséquences du facteur temporel de la citation est l'aspect unidirectionnel de ce graphe. Comme nous l'avons évoqué plusieurs fois, le graphe de citation est orienté. En effet, si l'article A_i émet une citation vers l'article A_j, cela n'implique pas que A_j cite en retour A_i. C'est d'ailleurs tout à fait impossible puisque A_j a été publié avant l'apparition de A_i. Le graphe de citation est donc unidirectionnel[5] : l'existence d'un arc $a_k = (n_i, n_j)$ rejette la possibilité d'un arc $a_{k'} = (n_j, n_i)$. L'existence de *circuit*, c'est-à-dire de chemin dont les extrémités coïncident, est alors impossible. Le graphe de citation est donc *orienté*, *unidirectionnel*, et *sans circuit* (fig. 2.4).

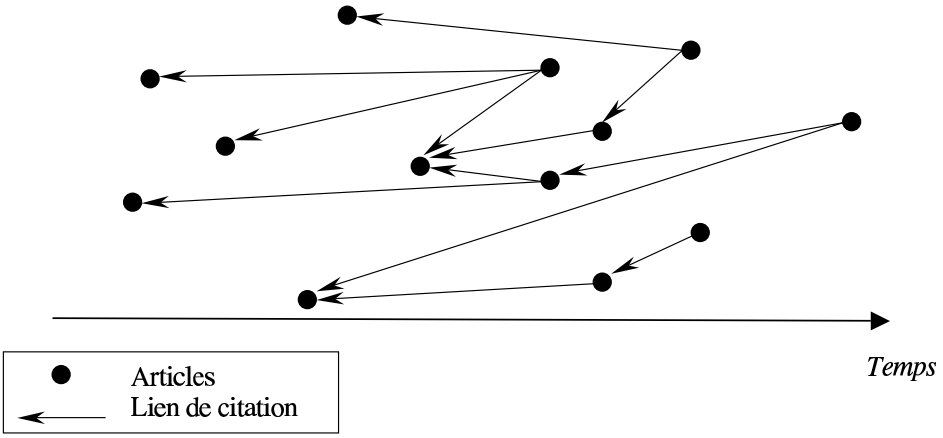

FIG. 2.4 – le graphe de citation : graphe orienté, unidirectionnel et sans circuit

2.2.2.2 Distribution des degrés des sommets

La distribution des degrés sortants (fig. 2.5), c'est-à-dire des citations émises par les articles est de type gaussienne assymétrique négative (mode<médiane<moyenne) avec un nombre moyen de références bibliographiques par article variant entre 20 et 50 [Zitt and Bassecoulard, 1998].

La distribution des degrés entrant, c'est-à-dire des citations reçues, suit une loi hyperbolique, lois que l'on rencontre souvent en sciences humaines et

[5] antisymétrique au sens mathématique.

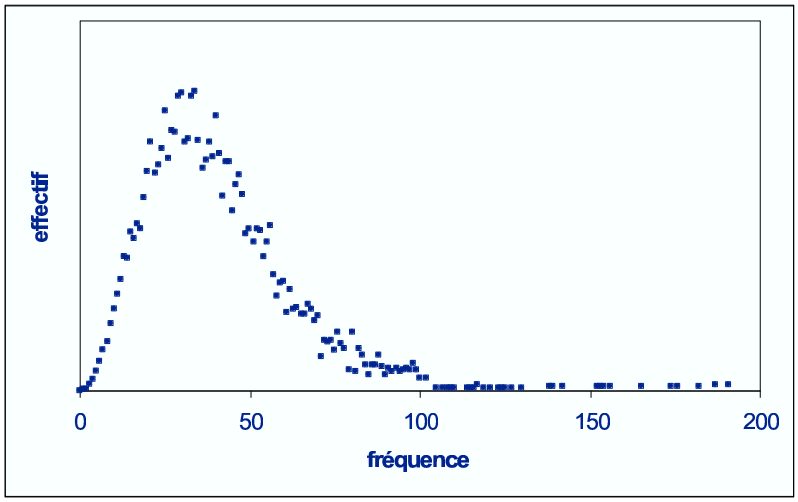

FIG. 2.5 – Distribution des citations émises (courbe réalisée à partir d'un corpus de 13.000 articles)

sociales. Trois lois hyperboliques sont particulièrement connues en bibliométrie. Il s'agit de :

- la loi de Zipf, loi caractérisant la distribution des mots dans les textes [Zipf, 1949],
- la loi de Lotka, loi de répartition des auteurs en fonction de leur nombre de publications [Lotka, 1926],
- la loi de Bradford, loi de répartition des revues en fonction de leur nombre d'articles [Bradford, 1934].

Les distributions hyperboliques sont caractérisées par un faible cœur et une forte dispersion (fig. 2.6). Le cœur représente un petit nombre d'éléments ou d'individus ayant une forte fréquence : peu d'articles sont très fréquemment cités, peu de mots sont très fréquents dans les textes, peu d'auteurs publient beaucoup. La dispersion caractérise un très grand nombre d'éléments ou d'individus ayant une faible fréquence : la majorité des articles reçoivent peu de citations, la plupart des termes n'apparaissent qu'une seule fois dans les textes, etc.

La distribution des citations (degrés entrants) est semblable à la répartition des mots dans les textes avec « quelques différences morphologiques mineures » [Zitt and Bassecoulard, 1998]. Elle respecte avec une pente plus faible que pour les mots la loi de Zipf. D'après Zipf si l'on dresse une table de l'ensemble des mots différents d'un texte quelconque, classés par ordre de fréquences décroissantes, on constate que la fréquence d'un mot est inversement proportionnelle à son rang, ou, autrement dit, que le produit de la fréquence (notée $g(r)$) de n'importe quel mot par son rang (noté r) est constant, ce que traduit la

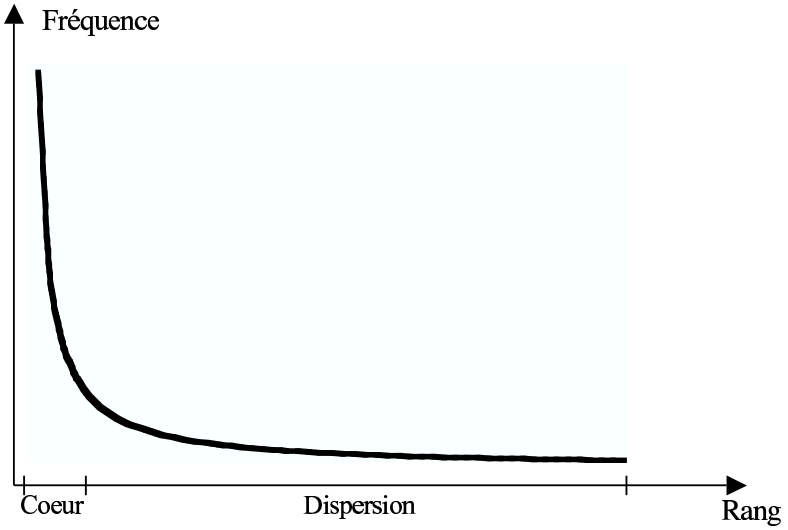

FIG. 2.6 – Cœur et dispersion des lois hyperboliques

formule :

$$g(r) = \frac{a}{r} \tag{2.10}$$

avec a une constante. La forme généralisée de Zipf s'écrit :

$$g(r) = \frac{a}{r^{\gamma}} \tag{2.11}$$

où γ tend vers 1 pour les mots. Les études empiriques montrent que γ est nettement plus faible pour les citations, proche de 0,4.

2.2.3 Les méthodes d'analyse du graphe de citation

Comme pour l'étude des réseaux sociaux, il existe différentes méthodes d'analyse du graphe de citations. Nous présenterons dans un premier temps deux familles d'indicateurs mesurant l'importance des sommets : les facteurs d'impacts et les facteurs d'influence ; puis nous présenterons les méthodes de structuration du graphe, permettant de détecter des sous-groupes cohésifs.

2.2.3.1 Les facteurs d'impacts et les facteurs d'influence

Le *facteur d'impact* est un des premiers indicateurs créé pour mesurer l'utilité des unités scientifiques. Sous sa forme généralisée, le facteur d'impact d'une unité scientifique pour la période T_1 se définit de la manière suivante [Ingwersen, 1998] :

$$IPF(T_1) = \frac{CIT_{T_2}(T_1)}{PUB(T_2)} \tag{2.12}$$

où

– $CIT_{T_2}(T_1)$ est le nombre de citations émises sur une période T_1 vers les items de l'unité scientifique publiés sur une période T_2 (avec T_2 antérieur à T_1)

– et $PUB(T_2)$ le nombre d'items de l'unité scientifique publiés sur la période T_2.

Dans l'exemple de la figure 2.7, l'ensemble des items de l'unité scientifique reçoit trois citations, et l'unité scientifique comporte trois items. Le facteur d'impact de cette unité est donc 1.

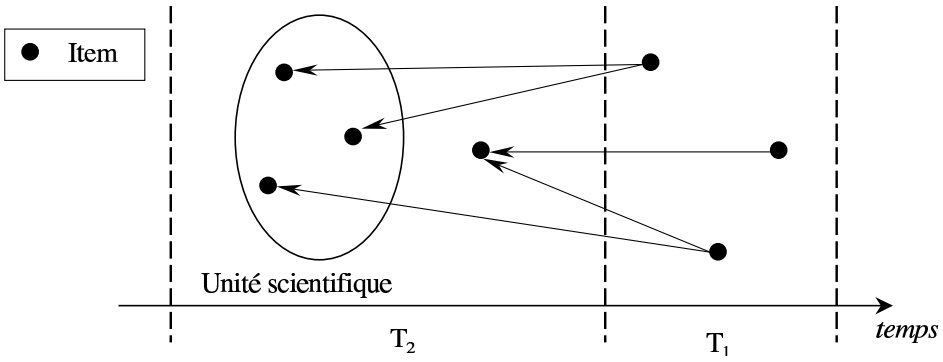

FIG. 2.7 – Calcul des facteurs d'impact

Le premier facteur d'impact proposé par Garfield [Garfield, 1972] et utilisé dans le *Journal citation reports (JCR)* de l'ISI, mesure chaque année l'audience des revues scientifiques. La période T_1 est une année donnée, prenons par exemple 2003. La période T_2 correspond aux deux années antérieures, ici 2001 et 2002. Le facteur d'impact d'une revue pour l'année 2003 est donc le nombre de total de citations faites pendant l'année 2003 aux articles de la revue publiés en 2001 et 2002, divisé par le nombre d'articles de la revue publiés en 2001 et 2002. Cet indicateur mesure donc la fréquence moyenne de citation des articles de ce journal, pour une période de deux ans. C'est un indice de mesure rétrospective de l'impact à relativement court terme.

Les facteurs d'impact sont proches de la mesure du prestige de degré évoqué dans la section 2.1.3.2. Construisons le sous-graphe de citations réduit aux périodes T_1 et T_2 (fig. 2.8). Dans ce sous-graphe, le facteur d'impact d'une unité scientifique appartenant à T_2 est la moyenne du prestige de degré (non normalisé) de chacun des sommets n_i, avec une limite importante : les liens existants entre les sommets de T_2 ne sont pas pris en compte (liens en pointillés sur la figure 2.8).

$$IPF(T_1) = \overline{P_D(n_i)} = \overline{d_e(n_i)} \tag{2.13}$$

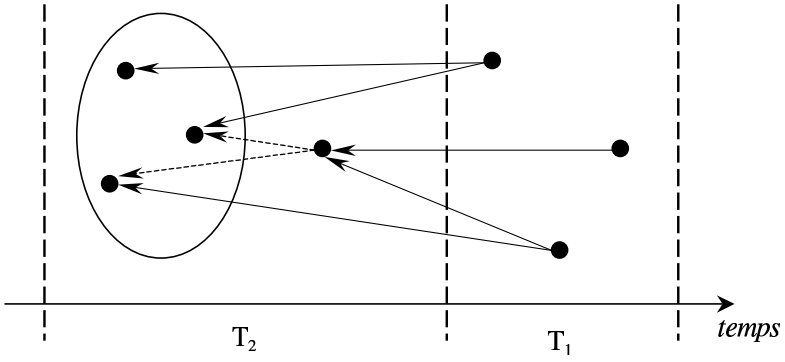

FIG. 2.8 – Sous-graphe de citation pour les périodes T_1 et T_2

En 1976, Pinski et Narin [Pinski and Narin, 1976] évoquent les limites des facteurs d'impacts.

- Premièrement, ceux-ci ne tiennent pas compte de la longueur moyenne des articles. Par exemple, les articles de synthèse sont souvent plus longs et couvrent plus largement un domaine scientifique que les articles de recherche. C'est pourquoi, ils sont fréquemment cités. Les revues proposant des articles de synthèse ont donc un meilleur facteur d'impact.
- Deuxièmement, les facteurs d'impacts ne prennent pas en considération les différentes pratiques de citations suivant les domaines de recherche.
- Enfin, pour le calcul de ces indicateurs toutes les citations ont la même valeur sans tenir compte de quelles revues elles proviennent.

Ces deux auteurs proposent donc un nouvel indicateur pour mesurer l'importance des revues, appelé *facteur d'influence*. Il prend en compte le fait que toutes les citations n'ont pas la même valeur : les citations issues des revues considérées comme importantes ont plus de valeur que les autres ; l'importance d'une revue étant elle-même définie par le nombre de citations reçues. Ce problème a reçu un formalisme mathématique que Egghe développe dans son ouvrage *Introduction to Informetrics* [Egghe and Rousseau, 1990].

Soit $A = (A_{ij})$ la matrice de citations de revues $n \times n$, où A_{ij} est le nombre de références bibliographiques issues d'articles publiés dans la revue i pointant vers des articles de la revue j. Le poids de l'influence w_i d'une référence bibliographique issue d'une revue i s'écrit comme suit :

$$w_i = \sum_{k=1}^{n} \frac{w_k A_{ki}}{S_i} \tag{2.14}$$

où $S_i = \sum_{j=1}^{n} A_{ij}$, c'est-à-dire le nombre total de références bibliographiques de la revue i, et w_k est le poids de l'influence d'une référence bibliographique issue d'une revue k. On obtient donc un système à n équations pouvant se résoudre

soit par des méthodes mathématiques exactes comme la méthode du pivot de Gauss, soit par des méthodes itératives qui calculent des suites convergeant vers le résultat. Egghe présente une résolution par un procédé itératif en posant à l'étape m,

$$w_i^{(m)} = \sum_{k=1}^{n} \frac{w_k^{(m-1)} A_{ki}}{S_i} \tag{2.15}$$

Il prend comme première estimation $w_i^{(1)}$ calculé pour une période donnée.

$$w_i^{(1)} = \frac{\text{nombre total de citations reçues par la revue } i}{\text{nombre total de références issues de la revue } i} \tag{2.16}$$

Les poids d'influence w_i obtenus mesurent l'influence des références bibliographiques. Pour avoir l'influence d'un article, il faut additionner les poids d'influence des références bibliographiques qui le citent. Pour obtenir l'influence d'une revue, on procède de même. On obtient (d'après la formule 2.14) :

$$\sum_{k=1}^{n} w_k A_{ki} = w_i S_i$$

Le facteur d'influence d'une revue i pour une année donnée se calcule de la manière suivante :

$$\frac{w_i S_i}{PUB(i)} \tag{2.17}$$

où $PUB(i)$ est le nombre d'article publiés dans la revue i pour l'année en question.

2.2.3.2 Les méthodes de structuration

L'un des objectifs de Price était de réaliser une carte de la science. Il voulait utiliser les résultats de la théorie des graphes pour les appliquer aux réseaux de publications scientifiques. Les méthodes de structuration, c'est-à-dire les méthodes permettant de détecter des groupes de publications cohésifs, telles que nous les avons présentées dans la section 2.1.4, ne sont pas directement applicables au graphe de citations. En effet, celui-ci est unidirectionnel et sans circuit, il ne comporte donc ni cliques, ni n-cliques, n-clans et n-clubs. C'est pourquoi les différentes méthodes de structuration de l'univers scientifique n'utilisent pas directement le graphe de citations mais des graphes associés. Des exemples de graphes associés sont les graphes d'unités scientifiques, comme les graphes d'auteurs ou les graphes d'institutions, graphes orientés, bidirectionnels et autorisant l'existence de circuits. L'étude des graphes de citations d'auteurs par exemple, permet de détecter des collèges invisibles[6].

[6]Selon Price, les collèges invisibles sont des groupes restreints d'auteurs, considérés comme les plus prolifiques d'un domaine de recherche. Bien que de nationalités différentes, leurs collaborations sont intenses. Ils sont les moteurs du domaine de recherche concerné et sont largement cités.

Nous allons présenter deux méthodes de structuration utilisées en bibliométrie citationniste.

1. La méthode du couplage bibliographique
   Kessler du *Massachusetts Institute of Technology* est le premier qui a essayé d'enrichir la méthode de l'analyse des citations. En 1963 [Kessler, 1963], il introduit la méthode du couplage bibliographique (*bibliographic coupling*). Son hypothèse est la suivante : si des documents possèdent des références bibliographiques identiques, il ont probablement une parité thématique. On dit que deux articles sont couplés s'ils partagent au moins une référence bibliographique commune. La force du couplage est déterminée par le nombre de références bibliographiques en commun.

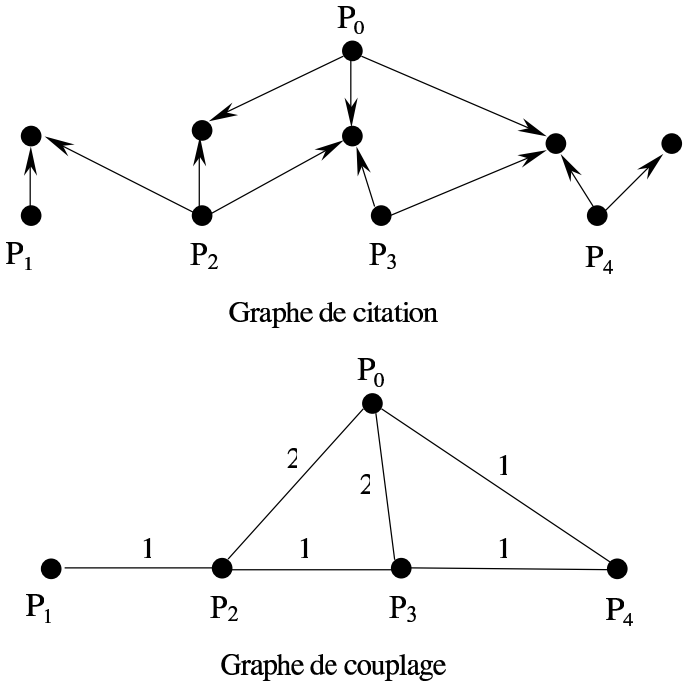

FIG. 2.9 – Construction d'un graphe de couplage à partir d'un graphe de citation

Cette méthode se base sur le graphe de couplage, où les nœuds sont les articles et les liens indiquent les relations valuées de couplage entre les articles (fig. 2.9). Kessler définit deux critères de couplage appelés A et B.
- Critère A (noté G_A) : Soit P_0 une publication scientifique. $G_A(P_0)$ est l'ensemble des articles couplés avec P_0, et $G_A(P_0; n)$ le sous-ensemble de $G_A(A_0)$ constitué des articles ayant exactement n références communes avec P_0.
- Critère B (noté G_B) : ensemble d'articles dans lequel chaque élément a au moins une référence commune avec chacun des autres éléments. Il s'agit des cliques dans le graphe couplage.

Dans l'exemple de la figure 2.9, nous avons :
$G_A(P_0) = \{P_2; P_3; P_4\}$ et $G_A(P_0; 2) = \{P_2; P_3\}$.
Les ensembles $\{P_0; P_2; P_3\}$ et $\{P_0; P_3; P_4\}$ vérifient le critère G_B.

Kessler voit cette méthode comme une aide à la recherche documentaire. En effet, le critère A permet d'élargir le nombre d'articles potentiellement intéressants pour un utilisateur. A partir d'un document pertinent, on peut facilement déterminer les documents couplés avec celui-ci. Le critère B correspond à la détection de cliques dans le graphe de couplage. Il permet de rapprocher thématiquement les documents d'un corpus. Au niveau théorique, une critique fondamentale a été évoquée [Martyn, 1964] : un article scientifique mêle souvent plusieurs idées, concepts ou méthodes, et peut être cité pour des intérêts totalement différents.

2. La méthode des co-citations
   La méthode des co-citations de documents a été mise au point indépendamment par deux chercheurs, Irina Marshakova [Marshakova, 1973], chercheur russe, et par l'américain Henry Small [Small, 1973]. Cette méthode est une véritable réalisation du projet de Price. Elle permet de créer des cartes relationnelles de documents, qui reflètent à la fois les liens sociologiques et thématiques d'un domaine de recherche. Elle repose sur l'idée évoquée par Price, selon laquelle la science est combinatoire. Chaque chercheur y ajoute sa brique. Si deux références bibliographiques de date quelconque apparaissent souvent ensemble, alors elles indiquent une combinaison intéressante entre les deux articles cités.

   Cette méthode vise la structuration d'un domaine de recherche pour une période restreinte (et généralement récente) que l'on veut étudier. Elle apparaît comme un renversement du couplage bibliographique. En effet, contrairement au couplage qui ne rapproche que les articles citants, cette méthode rapproche dans un premier temps les articles cités. Elle se base sur le graphe de co-citation, où les nœuds sont les articles et les liens indiquent les relations valuées de co-citation, c'est-à-dire le nombre de fois que les articles sont cités ensemble par d'autres articles. La figure 2.10 montre par exemple que les publications P_2 et P_3 ont un lien de co-citation de force 2, en effet ces deux publications sont citées à la fois par P_4 et P_5.

   L'analyse des co-citations met en relation deux ensembles :
   – l'ensemble des publications appartenant au domaine et à la période d'étude, les *éléments citants*,
   – et l'ensemble des publications citées par celles-ci, les *éléments cités*.

   Pour que cette analyse soit significative, le corpus des éléments citants doit être un ensemble homogène dans lequel les pratiques de citations sont similaires. Cette méthode comporte deux étapes :

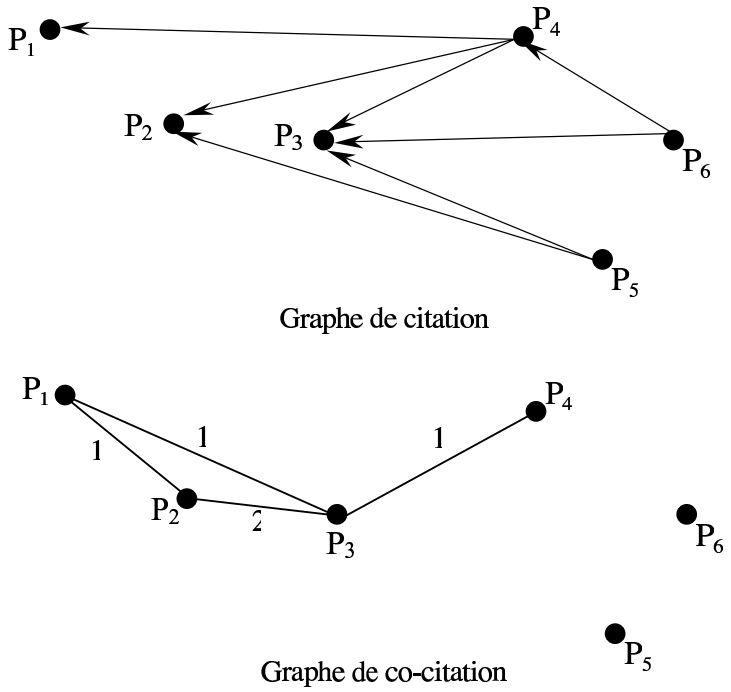

FIG. 2.10 – Construction d'un graphe de co-citation à partir d'un graphe de citation

- Le regroupement des éléments les plus cités en agrégats (*clusters* en anglais). Ce regroupement s'effectue après avoir évalué la proximité entre les articles. Sont considérés comme proches, les couples d'articles ayant une fréquence de co-citations élevée par rapport à leurs fréquences de citation respectives. L'application de méthodes issues de l'analyse de données permet de regrouper ensemble les articles les plus proches[7] .
- L'affectation des documents citants à ces clusters. Un article est assigné à un cluster avec un seuil x, s'il a au moins x références à des articles contenus dans le cluster.

L'intérêt de cette méthode réside dans son aspect dynamique. La structuration de la littérature citée permet de mettre en relation des publications plus ou moins anciennes. Par contre, la structuration des éléments citant rapprochent les publications de la période d'étude. Elle donne le reflet de l'état de la science pour la période étudiée. Elle permet d'identifier les *fronts de recherche*[8], c'est-à-dire les différentes orientations du domaine.

[7]Une méthode de regroupement sera présentée au cours du chapitre 4, section 4.5.

[8]Price divise la littérature scientifique en deux groupes : l'archive qui regroupe des documents souvent anciens, précurseurs dans un domaine de recherche (articles de référence) et les fronts de recherche. Un front de recherche est un ensemble de documents plus récents très fortement reliés entre eux, qui témoignent de l'émergence d'une nouvelle thématique dans ce domaine.

La méthode originelle des co-citations a été très critiquée, notamment par des sociologues [Hicks, 1987]. Une de ses faiblesses majeures est le découpage très partiel de la littérature citante. Dans l'expérience décrite par l'équipe de l'ISI[9] [Small and Sweeney, 1985], pour des contraintes techniques et des choix théoriques infométriques, seuls 1% des documents cités participent à la structuration du domaine. Dans les expériences de co-citations réalisées avec cet algorithme, on assigne dans le meilleur des cas 20 à 30 % du corpus à des fronts de recherche. Les chercheurs de l'INRA, du laboratoire LERECO [Zitt and Bassecoulard, 1996] ont montré qu'en utilisant un algorithme de classification efficace (le lien moyen aménagé) et en enrichissant les cœurs de co-citation par des articles plus faiblement cités, on obtient d'excellents taux de rappel, variables toutefois selon les domaines. Environ 99% de la science citante peut être affectée à des clusters.

La méthode des co-citations d'article a été adaptée aux réseaux de l'unité scientifique « auteur ». La méthode des co-citations d'auteurs a été introduite en 1981 [White and Griffith, 1981]. La structuration obtenue par cette méthode est moins précise du point du vue thématique que celle réalisée par la méthode des co-citations d'articles. Cependant elle met davantage en évidence les liens sociologiques entre les chercheurs. Sur les cartes, deux auteurs sont proches si à la fois ils travaillent sur le même thème de recherche et collaborent ensemble.

2.2.4 Les limites de l'analyse des citations

Dans l'introduction de cette section (2.2), nous avons brièvement présenté les deux approches possibles pour analyser les univers scienfiques : l'approche lexicale et l'approche citationniste. Comme nous l'avons évoqué, la première approche se heurte aux difficultés bien connues lorsque l'on travaille sur la langue. Un certain nombre de travaux critiques s'interrogent aussi sur les limites de l'analyse des citations. Certaines de ces critiques sont directement liées à la mise en œuvre des méthodes, d'autres plus profondes reposent sur les fondements théoriques de cette analyse.

2.2.4.1 Les limites liées à la mise en œuvre

- Couverture des bases de l'ISI
  En pratique, la quasi-totalité des expériences « citationnistes » utilisent les sources de l'ISI. Ces bases de données représentent-elles correctement la masse des publications scientifiques ? Certes, la volonté

[9] L'expérience de l'ISI, de 1985 porte sur 500.000 articles citants engendrant 3,9 millions d'articles cités. Seuls les 44.000 documents les plus cités sont utilisés pour la classification et au mieux 66% sont classés dans des agrégats.

de l'éditeur est d'indexer toutes les revues considérées comme importantes de chaque domaine de recherche. Pourtant, nous savons qu'il existe un biais favorable envers les revues américaines et anglophones qui sont davantage représentées [10]. Cet avantage augmente la visibilité des travaux américains, qui d'ailleurs est amplifiée par le fait que les auteurs américains ne citent pratiquement que des travaux américains [Cronin, 1981][11].
De plus, l'incomplétude des bases de données de l'ISI, en particulier pour les revues nationales, pose le problème de la validité des études bibliométriques à l'échelle micro.

– Homographes et Synonymes
On retrouve ici des limites analogues à celles rencontrées dans l'approche lexicale (homonymes, synonymes). Les *homographes* sont des auteurs ayant le même nom et les mêmes initiales. Ce problème est courant pour les noms japonais ou chinois. Pour les différencier il est alors nécessaire de connaître leur affiliation institutionelle. Les *synonymes* sont des auteurs connus sous plusieurs formes. Par exemple, les auteurs dont le nombre d'initiales varie, les femmes mariées empruntant le nom de leur mari, ou l'accolant à leur nom de jeune fille, etc.

2.2.4.2 Les limites de la méthode

– Les motivations de citations
L'analyse des citations repose sur l'hypothèse de la science combinatoire évoquée initialement par Price. Depuis quarante ans, de nombreux auteurs et en particulier des sociologues cherchent à comprendre quelles sont les véritables raisons qui amènent un scientifique à citer le travail d'un autre [Case and Higgins, 2000]. Deux écoles de pensée coexistent. Pour Merton, la science est un système normé dans lequel la citation est avant tout la reconnaissance d'une dette intellectuelle (sauf dans le cas de la citation critique). D'après lui, les auteurs d'un article mettent en avant leur propre travail, mais reconnaissent aussi le mérite d'autres travaux qui ont pu les inspirer. D'autres auteurs, comme Bruno Latour ou Leopold, évoquent le pouvoir de persuasion de la citation. Selon eux, la citation sert largement les intérêts des auteurs qui l'utilisent (les auteurs citants). En effet, elle peut être employée dans un souci de crédibilité, comme un moyen pour s'immuniser contre la critique [Gilbert, 1977]. Les études empiriques montrent que l'une et l'autre de ces motivations apparaissent au sein d'un même document. Latour reconnaît lui aussi l'aspect persuasif de la citation. Lorsqu'un auteur a le choix de citer pour un même champ d'investigation, les travaux d'un auteur réputé

[10] A titre d'exemple le SSCI dépouille intégralement les articles de 1794 revues en sciences sociales, seules une trentaine sont en français.

[11] Dans l'étude réalisée par Cronin en 1981, 95% des citations émises par des auteurs américains pointaient vers des travaux américains. Ce qui était bien plus que la part des travaux américains dans l'indicateur de la production scientifique mondiale.

ou ceux d'un auteur moins connu, on observe que c'est l'auteur réputé qui est cité.

D'autre part, sans prendre en considération l'objectif de citation (rendre hommage ou persuader) il existe plusieurs raisons « spécifiques » de citer un document. Plusieurs auteurs comme Garfield listent ces raisons. Ce dernier en cite quinze [Garfield, 1965] parmi lesquelles : rendre hommage aux pionniers ; montrer l'intérêt de travaux antérieurs ; identifier des méthodes, des protocoles ; corriger son propre travail ; corriger le travail des autres ; critiquer des travaux antérieurs ; etc. D'autres auteurs s'intéressent aux liens typés entre les documents citants et cités, ou examinent le contexte dans lequel la citation a été faite. De toutes ces analyses, l'on s'apperçoit que l'acte de citation, même réalisé par un auteur identique, est effectué dans des conditions non-équivalentes : un article peut être référencé pour son intérêt général ou pour un intérêt partiel. Il est important de noter aussi que les scientifiques ne citent pas tous les travaux qui les ont influencés, et que parfois ils ont recours à la citation implicite.

Nous insistons sur le fait que les différentes méthodes d'analyse des citations présentées ne distinguent pas ces différentes motivations. Pour bien comprendre l'évolution d'un domaine de recherche, des analyses qualitatives sont aussi nécessaires.

– L'inertie du système de citation
Un document, aussi intéressant soit-il, ne peut être cité dès sa parution. En effet, même s'il est repéré immédiatement et référencé dans un article à soumettre, la procédure de publication de l'article le citant peut s'étendre sur une période de plusieurs mois. Contrairement à l'approche lexicale qui permet une analyse instantanée de la science, dans l'approche citationniste les délais de publications constituent un frein pour une analyse de la science « en temps réel ».

2.3 Le graphe du Web

Le Web, comme tout système hypertexte se présente sous la forme d'un graphe orienté où les nœuds sont des éléments d'information et les arcs les liens hypertextes permettant la navigation.

1. Les éléments d'information
Dans les systèmes hypertextes, les éléments d'information représentent généralement des documents élémentaires dont le contenu exprime un nombre limité d'idées [Balpe et al., 1996]. Ainsi, ils sont autonomes d'un point de vue sémantique. Sur le Web les nœuds sont des fichiers d'information au format HTML (pages web) qui contiennent eux-mêmes des liens

vers des fichiers de nature hétérogène comme des images, du son, de la vidéo et de plus en plus souvent des documents au format PDF ou PS. Les nœuds d'information sont localisables grâce à leurs adresses URL (*Uniform Resource Locator*). La forme normalisée d'une adresse URL s'écrit de la manière suivante :

protocole ://machine.nom-de-domaine/adresse-fichier ;paramétres ?requete#argument[12]

2. Les liens hypertextes
   Les liens hypertextes permettent la navigation de page en page. Ils sont composés de deux parties : une ancre et une cible. L'ancre est le point de départ du lien qui peut être activé par un clic de souris. Cela peut être un morceau de texte, une icône, une image. La cible donne le fichier de destination identifié par une adresse URL. Dans le langage HTML, la balise <A> permet d'insérer des ancres et son attribut HREF donne la destination du lien :
   <A HREF="URL pointée" (cible)> texte ou image apparaissant à l'écran (ancre) </A>.

Communément ce graphe est appelé *graphe de citation* par analogie avec la bibliométrie. Toutefois, nous verrons dans la section 2.3.2 les limites de ce rapprochement. D'autre part, ce graphe est binaire. Il exprime la présence ou non d'une relation orientée d'une page vers une autre. Autrement dit, le fait que la page soit reliée plusieurs fois par une autre est considérée comme une information redondante et n'est pas prise en compte.

2.3.1 Topologie et mesure du graphe

De nombreux chercheurs s'intéressent à la taille et la topologie du Web avec de nombreuses applications possibles comme l'amélioration des algorithmes de crawling ou de la recherche d'information ou encore l'étude des phénomènes sociologiques.

2.3.1.1 Estimation de la taille du Web

Lawrence and Lee Giles du NEC Research Institute ont proposé en 1998 [Lawrence and Giles, 1998] une méthode pour estimer la taille du Web indexable[13].

[12] Le protocole de transfert de fichier utilisé sur le Web est HTTP (HyperText Transfert Protocol)

[13] A l'époque le Web indexable concernait l'ensemble des pages statiques (pages non générées par des requêtes) au format HTML accessibles sans restriction (mot de passe), en d'autres termes les pages pouvant être indexées par les moteurs. Une telle définition semble désuète aujourd'hui puisque le moteur Google indexe désormais des documents sous différents formats comme les formats .pdf, .doc

Cette méthode croise les réponses obtenues par plusieurs moteurs pour différentes requêtes (fig. 2.11). Soit N le nombre pages du Web indexable. Pour une requête donnée, si le moteur A rend n_a réponses et le moteur B en rend n_b et si n_0 est le nombre de réponses communes, alors la quantité $p_a = n_0/n_b$ est une estimation de la portion du web indexable prise en compte par A. Une estimation N du Web indexable s'obtient par le produit $\frac{N_a}{p_a}$, où N_a est le nombre de pages indexées par le moteur A.

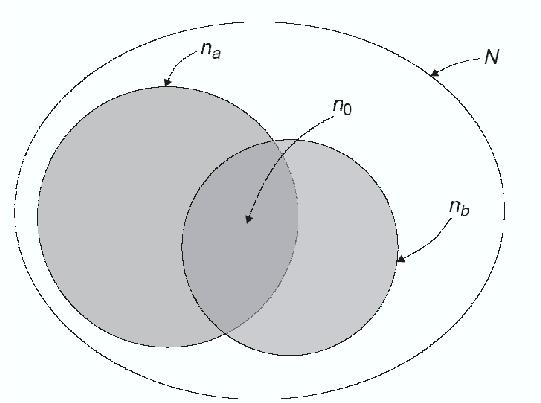

FIG. 2.11 – Estimation de la taille du Web selon Lawrence and Lee Giles (1998)

Cette équipe estimait la taille du Web indexable en décembre 1997 à 320 millions de pages. Une autre étude menée en février 1999 par cette même équipe donne la taille du Web indexable à 800 millions de pages et considérait que le meilleur moteur de recherche (Northern Light) couvrait au mieux 16% du Web [Lawrence and Giles, 1999]. En combinant les meilleurs moteurs on pouvait atteindre une couverture de 42%. Cependant, cette étude montrait aussi que la mise à jour des index décroît en fonction de la couverture du Web. Ces résultats sont légèrement contestés par l'équipe du Compaq Systems Research Center [Bharat and Broder, 1998] qui pense que ces chiffres sont surévalués. En effet, Lawrence et Giles ne tiendraient pas compte dans leur méthode de la duplication des pages, phénomène pourtant très présent sur la toile (site mirroirs, etc.). Bharat et al. avancent pour mars 1998 le chiffre de 275 millions de pages web statiques distinctes. Différentes sources l'estiment actuellement à 3 milliards de pages, et Google indexerait plus de 2 millards de documents (pages web et autres fichiers).

Par contre, la taille du Web invisible est difficilement identifiable. Si l'on définit le Web visible comme le Web accessible par les moteurs de recherche, le Web invisible est par opposition l'ensemble des pages non indexées pour diverses

raisons : pages nouvelles et inconnues des moteurs, pages dynamiques, pages aux accès limités par mot de passe, etc. Nous savons que la taille du Web invisible est nettement plus importante que celle du Web visible à cause des nombreux serveurs de bases de données disponibles. A titre d'exemple le serveur Dialog[14] est capable à lui seul de générer 5 milliards de pages.

2.3.1.2 Distribution des degrés des sommets

Les distributions des degrés entrants d_e et sortants d_s suivent des lois hyperboliques comparables à la répartitions des citations reçues en bibliométrie (section 2.2.2.2, page 29). Ces deux distributions peuvent s'écrire sous la forme de lois de probabilité puissance, comme le montrent les deux équations suivantes :

$$P(d_e(n_i) = k) = \frac{C_1}{k^{\lambda_e}} \tag{2.18}$$

$$P(d_s(n_i) = k) = \frac{C_2}{k^{\lambda_s}} \tag{2.19}$$

où C_1, C_2 et λ_e, λ_s sont des constantes. Trois études menées sur des corpus de tailles différentes [Kumar et al., 1999][15] [Albert et al., 1999][16] [Broder et al., 2000][17] tentent d'évaluer les facteurs λ_e, λ_s. Toutes les trois montrent que $\lambda_e = 2,1$, c'est-à-dire que la probabilité d'obtenir une page citée k fois est proportionnelle à $1/k^{2,1}$. Les valeurs obtenues pour λ_s sont un peu plus élevées : 2,45 [Albert et al., 1999] et 2,72 [Broder et al., 2000], ce qui indique que le nombre de liens émis par page est plus dispersé. La figure 2.12 donne les résultats obtenus par Broder et al.

2.3.1.3 Connectivité et diamètre

Albert et al. [Albert et al., 1999] s'intéressent aux propriétés topologiques du graphe et essayent plus particulièrement de déterminer son diamètre. Dans cette optique, ils modélisent le Web en construisant de manière aléatoire un graphe orienté composé de N nœuds qui vérifie les distributions hyperboliques présentées dans la section précédente (avec $\lambda_e = 2,1$ et $\lambda_s = 2,45$). Ils calculent les distances d entre les sommets, c'est-à-dire la longueur du plus court chemin entre deux nœuds. Ils montrent que la distance moyenne entre deux nœuds est une fonction logarithme de l'ordre du graphe :

$$Moy(d) = 0.35 + 2,06 \log(N). \tag{2.20}$$

[14] `http://www.dialog.com`

[15] Cette étude utilise un corpus de 40 millions de pages.

[16] Cette étude se base sur les 325 729 pages du domaine de l'université Notre Dame "nd.edu" qui comportent presque 1.5 millions de liens.

[17] Le corpus de cette étude contient 200 millions pages et 1.5 milliards de liens obtenus par les crawls du moteur altavista.

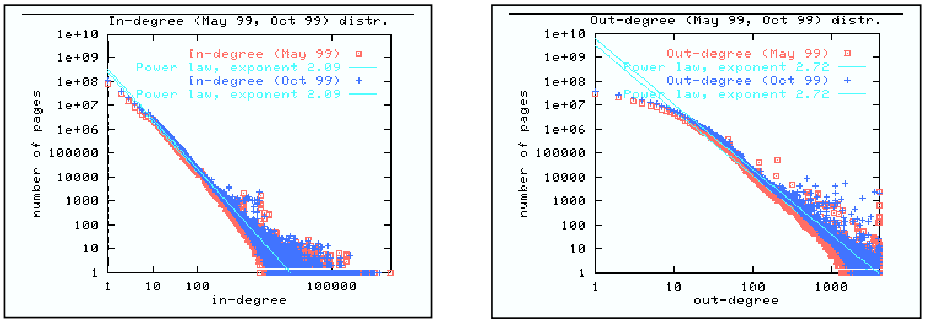

FIG. 2.12 – Distribution des degrés entrant et sortant d'après Broder et al. (2000)

Au moment de l'expérience la taille du Web était estimée à 800 millions de pages et les auteurs évaluent la distance moyenne entre deux pages par l'équation ci-dessus à 19[18]. En théorie, on appelle *diamètre* d'un graphe la longueur de la plus longue chaîne de ce graphe, c'est-à-dire la distance maximum entre deux sommets. Albert et al. introduisent la notion de diamètre moyen : ils remarquent que pour un N donné, d suit une distribution gaussienne, et selon les auteurs peut être interprété comme le diamètre moyen du graphe.

Cette modélisation présente le Web comme un réseau « petit monde », c'est-à-dire un univers de faible diamètre et fortement interconnecté. La théorie du petit monde ou « des six degrés de séparation » est une notion provenant de l'analyse des réseaux sociaux qui montre qu'il ne faut pas plus de six intermédiaires pour rejoindre n'importe qui dans le monde.

L'expérience menée par Broder et al. [Broder et al., 2000] sur 200 millions de pages web (soit un quart du web de l'époque) remet en cause ces résultats. Une expérience à si grande échelle a pu être réalisée grâce à l'utilisation du *Connectivity Server 2* développé par le Compaq Systems Research Center [Bharat et al., 1998]. Outre la distribution des degrés des sommets, cette expérience se penche sur la connectivité du Web en identifiant les *composantes connexes* de celui-ci. Une composante connexe est un ensemble de sommets dans lequel chaque paire est reliée par une chaîne et dont il n'existe pas d'autres sommets adjacents à ces sommets (sous-graphe induit maximal connexe). L'originalité de ce travail est d'étudier à la fois le graphe du Web (graphe orienté), et le graphe obtenu en enlevant les orientations de ce dernier. On distingue alors les *composantes fortement connexes (CFC)*, qui sont des composantes connexes en tenant compte des relations orientées du graphe, et les *composantes faiblement connexes (CfC)*, composantes connexes dans le graphe déchu de ses orientations.

[18] Ce qui signifie qu'une page peut être atteinte à partir d'une autre en cliquant en moyenne 19 fois.

Dans le cas du graphe non orienté, les auteurs découvrent une composante faiblement connexe « géante » comportant 91% des noeuds de graphe (186 millions de pages). D'autres composantes de taille plus faible sont repérées. La distribution de leurs tailles suit d'ailleurs une loi hyperbolique. Dans le cas du graphe orienté, il existe aussi une composante géante, mais de taille moindre (56 millions de nœuds, soit 28% du corpus).

Les auteurs réalisent une étude complémentaire utilisant une technique d'exploration des graphes nommée *recherche en largueur*[19] (*breadth-first search* en anglais). Cette étude permet de déterminer le taux de pages pouvant être atteintes à partir d'un ensemble de pages formé de manière aléatoire. Les résultats obtenus permettent de déduire la fameuse forme de nœud papillon du Web (fig. 2.13)

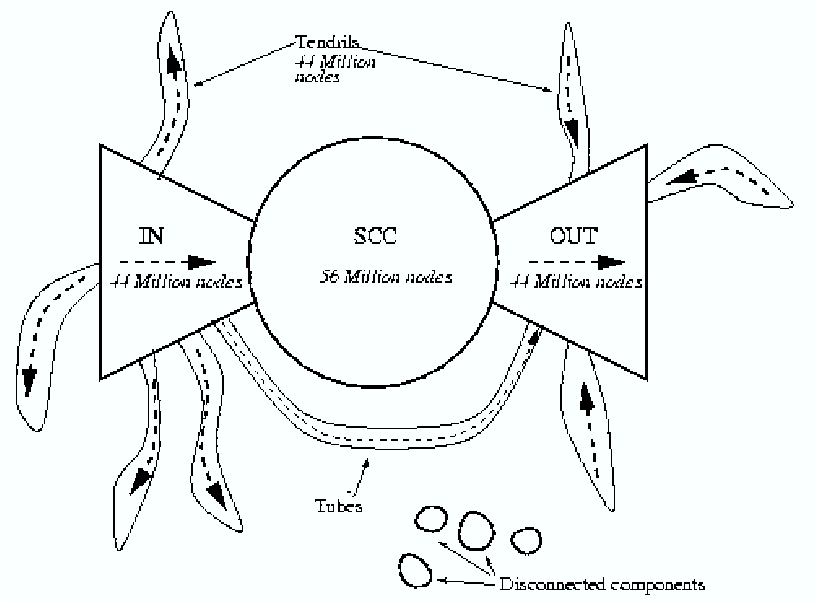

FIG. 2.13 – La théorie du nœud papillon selon Broder et al. (2000)

Selon les auteurs, le Web comporte :

– un cœur, CFC géante composée de 56 millions de pages (noté SCC sur la figure 2.13 pour *Strongly Connected Component*) dans laquelle toutes les paires de pages sont reliées par un chemin ; son diamètre est de 28 liens (selon la définition théorique du diamètre énoncée précédemment),

[19] La recherche en largeur permet d'explorer les graphes par couche. En partant d'un sommet u, la première couche regroupe tous les nœuds pouvant être atteints par un arc issu des u. La couche k regroupe les nœuds pouvant être atteints par un arc issu des nœuds de la couche $k-1$.

- un ensemble de pages noté *IN*, contenant presque 44 millions de pages qui peuvent atteindre de manière directe ou indirecte les pages du coeur,
- un ensemble de pages noté *OUT*, contenant aussi 44 millions de pages qui peuvent être atteintes à partir des pages du cœur,
- des *tendrils*, ensemble de pages ne pouvant ni atteindre le cœur ni être atteintes à partir de celui-ci (composé de 44 millions de pages aussi),
- des éléments complètement déconnectés (16 millions de pages environ).

Cette expérience montre que le Web est nettement moins interconnecté que ne l'avait prédit Albert et al. et que celui-ci a une structure plus subtile qu'un réseau « petit monde ». Le diamètre total du graphe est évalué à 500 et la probabilité pour que deux pages soient en relation directe ou indirecte est de 24% seulement. Par contre lorsque qu'il exite un chemin entre deux pages sa longueur moyenne est évaluée à 16, et seulement à 6 dans le graphe non orienté.

2.3.2 La Webométrie

La Webométrie est un nouvel axe de recherche qui s'intéresse à la fois :

1. aux bouleversements causés par l'apparition du Web dans le domaine de la recherche scientifique,
2. et à l'application des techniques bibliométriques dans ce nouvel univers.

1. Le Web et les autres services de l'Internet sont une aubaine pour les bibliomètres, car ils offrent de nouvelles sources d'information sur support numérique liées à l'activité scientifique (littérature grise, forums, etc.) différentes des traditionnelles bases de données d'articles. Cependant, l'utilisation du Web par les chercheurs remet aussi en cause certains des traditionnels indicateurs scientométriques. En effet, peu à peu les comportements de publication changent. Le Web permet de nouveaux environnements d'édition aux accès parfois gratuit [Cronin, 2001], et la publication sur le Web devient alors une alternative au système traditionnel. Cronin nous donne comme exemple le site web *PubMed Central (PMC)*[20] hébergeant des articles scientifiques dans le domaine biomédical. Quant à nous, nous connaissons particulièrement bien le moteur CiteSeer[21], réservoir d'articles en informatique. Lawrence, son créateur remarque d'ailleurs que les articles accessibles gratuitement en ligne sont largement plus cités [Lawrence, 2001] que les autres. Finalement n'est-on pas plus visible en publiant sur le Web que dans des revues à haut facteur d'impact ?

2. L'application des techniques bibliométriques au Web se fait dans deux optiques différentes :

- l'évaluation des différentes groupes présents sur la toile,

[20] `http://www.pubmedcentral.nih.gov`
[21] `http://citeseer.ist.psu.edu/cs`

– l'appréhension de cet univers, en vue d'améliorer son accès.

L'évaluation des différents groupes sociaux présents sur la Toile intéresse surtout les bibliométres et scientométres qui essayent de transposer leurs savoirs et expertises [Rostaing et al., 1999]. Traditionnellement l'évaluation générale des pays et des aires régionales se fait à partir de différents indicateurs démographiques et économiques, alors que les indicateurs bibliométriques se penchent plus spécialement sur l'évaluation de la production scientifique. Le Web étant un espace de publication universel, les indicateurs bibliométriques transposés dans cet espace peuvent venir compléter et renforcer les indicateurs généralistes. Au niveau macro, l'évaluation des pays, ou au niveau micro, l'évaluation des sites Web, peut s'effectuer par l'application des méthodes bibliométriques. Ainsi, Rousseau [Rousseau, 1997] s'intéresse en 1997 à la distribution des extensions de noms de domaines (représentant parfois des pays comme l'extension .fr pour la France) pour le champ de la bibliométrie et de la scientométrie. Ingwersen [Ingwersen, 1998] transpose la notion de facteurs d'impact et crée les *WIF (Web Impact factor)*.

L'orientation qui nous intéresse plus spécialement dans le cadre de cette thèse est l'utilisation des méthodes bibliométriques pour améliorer la recherche d'information. Curieusement, cette voie est particulièrement investie par la communauté des informaticiens qui transposent les méthodes basées sur l'analyse des citations. Le graphe du Web devient analogue au graphe de citation : les pages web jouent le rôle des articles et les liens hypertextes celui des références bibliographiques. Que penser d'une telle analogie ? Eghhe [Egghe, 2000] et Prime et al. [Prime et al., 2002b] nous en donnent quelques limites :

– Une différence majeure entre un article scientifique et une page web réside dans la volatilité et la possibilité de mise à jour de la page web. Ceci a pour conséquence une forme différente du graphe du Web par rapport au graphe de citation traditionnel. En effet, alors que deux articles scientifiques ne peuvent avoir de citation réciproque, il est tout à fait envisageable que deux pages se citent mutuellement. L'aspect unidirectionnel du graphe de citation disparaît complètement pour le graphe du Web, ce qui permet d'ailleurs l'application plus large des différentes méthodes provenant de l'analyse des réseaux sociaux. Le problème de la mise à jour pose aussi la question de la pertinence de la citation. Comment être sûr que le contenu d'une page citée par un auteur n'a pas complètement changé ou même que cette page n'a pas disparu. Pour remédier à ces problèmes, un groupe d'éditeurs américains proposent de donner à chaque document électronique un identifiant unique (appelé DOI pour *Digital Object Indentifier*) analogue aux ISBN et ISSN. La correspondance entre les DOI et les adresses URL se fait par l'intermédiaire d'un système centralisé géré par l'*International DOI foundation*[22] créé en 1998. Ce système ne concerne pour l'instant que les publications des éditeurs commerciaux.

[22] `http://www.doi.org/`

- La duplication des pages est une opération courante sur le Web. Elle est généralement effectuée pour permettre un plus rapide accès aux ressources. Certains serveurs très volumineux et souvent consultés évitent les encombrements en proposant plusieurs copies de leurs sites en différents points de la planète. On parle alors de sites miroirs. Cependant, ces reproductions génèrent la réplication des liens hypertextes, ce qui constitue une limite importante pour la transposition des indicateurs basés sur la citation. Les URN (Uniform Resource Name) donnant un nom unique et permanent pour désigner une ressource sont censées pallier à la fois le problème de la duplication des pages et celui de l'instabilité des adresses URL. Cependant, elles sont trop peu utilisées.
- Dans cette analogie le lien hypertexte matérialise une citation. Pourtant nous savons aussi qu'il est très souvent utilisé à d'autres fins comme celle de la navigation au sein d'un même site. De plus, nous ne pouvons ignorer la présence de liens publicitaires sur la Toile. Différents chercheurs s'interrogent sur la manière de typer les liens, mais une fois de plus, une telle tâche ne peut pas être imposée aux auteurs.

Nous présenterons dans la section suivante les méthodes de recherche d'information utilisant la structure du graphe, et montrerons comment elles se rapprochent de la bibliométrie et des réseaux sociaux.

2.3.3 Analyse du graphe et applications dans le cadre de la RI

Comme en Bibliométrie ou dans le domaine des réseaux sociaux, les techniques d'analyse du graphe du Web suivent deux orientations possibles :

- l'orientation « individuelle » permet d'attribuer des scores aux pages web dans l'optique de les classer. Les deux applications les plus connues sont le *Page Rank* et l'agorithme *HITS*.
- l'orientation « collective » permet d'organiser les ressources et de découvrir des communautés d'intérêt.

2.3.3.1 Un algorithme de classement des pages : le *Page Rank*

Le *Page Rank (PR)* est l'algorithme de classement de pages implémenté dans le moteur de recherche *Google*[23]. Devant l'efficience non concluante des moteurs dits de « première génération » qui s'appuient sur des techniques lexicales, Google propose de prendre en compte la dimension structurelle du Web [Brin and Page, 1998]. L'algorithme *Page Rank* repose sur l'hypothèse empruntée à l'analyse des citations, selon laquelle une page pointée par une autre reçoit une sorte de reconnaissance, d'approbation de la part de l'auteur de celle-ci.

[23] `http://www.google.com/`

Ainsi, une page fréquemment citée, donc populaire, est considérée comme plus importante et sera rendue en priorité à l'utilisateur. Les opposants de cette méthode s'interrogent sur la relation entre la popularité et la pertinence d'une page. Ses défenseurs expliquent que le *Page Rank* d'une page, c'est-à-dire la valeur calculée par l'algorithme PR pour cette page, est la probabilité pour qu'un utilisateur la visite en naviguant sur la toile de façon aléatoire de lien en lien (en ne revenant jamais en arrière) et que de cette manière il traduit une certaine réalité du Web. Les *Page Rank* de chaque page sont attribués par un algorithme itératif. Le PR d'une page A se calcule selon la formule suivante :

$$PR(A) = (1-d)(\frac{PR(T_1)}{d_s(T_1)} + ... + \frac{PR(T_n)}{d_s(T_n)}) \tag{2.21}$$

où

- $\{T_1; ...; T_i; ...; T_n\}$ est l'ensemble des pages qui citent A,
- $PR(T_i)$ le *Page Rank* de la page T_i,
- $d_s(T_i)$ est le degré sortant de la page T_i c'est-à-dire son nombre de références émises,
- le paramètre d est un facteur d'amortissement qui varie entre 0 et 1 permettant de faire converger l'algorithme plus rapidement. Il est fixé à $d = 0.85$ par les auteurs.

Cette équation traduit qu'une page A a un bon *Page Rank* si :

1. elle est énormément citée ($n = d_e(A)$ élevé)
2. elle est citée par des pages T_i ayant à la fois un bon PR et peu de références ($d_s(T_i)$ faible).

On retrouve ici non seulement la notion de prestige évoquée au paragraphe 2.1.3.2 (page 21), mais surtout les hypothèses de base des facteurs d'influences (section 2.2.3.1, page 33) à savoir : (i) le fait que toutes les citations n'ont pas la même valeur, (ii) l'importance (ici la popularité) d'une page dépend de l'importance des citations reçues. Le Page Rank n'est rien d'autre que l'application des facteurs d'influence adaptée au graphe du Web.

La force de Google réside en partie dans l'aquisition du graphe du Web et de sa mise à jour, et dans la rapidité de l'algorithme calculant les PR[24]. Le calcul des PR est effectué pour l'ensemble des pages indexées et est indépendant des requêtes.

2.3.3.2 Pages références et pages pivots : l'algorithme HITS

Comme S. Brin et L. Page, l'équipe du projet Clever [Gibson et al., 1998] entreprend la création d'un système de recherche d'information utilisant la struc-

[24]Les PageRank de 26 millions de pages Web peuvent être calculés en quelques heures sur un poste de travail de puissance moyenne.

ture du Web. Son objectif est de repérer parmi les pages réponses obtenues avec un moteur classique (Alta vista), celles qui font « références » (appelées *pages références*, *authorities* en anglais). et celles qui orientent l'utilisateur vers des pages références (appelées *pages pivots*, *hubs* en anglais). Prenons l'exemple d'un utilisateur cherchant des renseignements sur la société X. En exécutant la requête « société X » avec un moteur classique, l'ensemble des pages retrouvées contiendra à la fois des documents décrivant cette société X (homepage de la société par exemple) et des pages la mentionnant sans la décrire réellement. Les pages références décrivent ou discutent d'un sujet déterminé tandis que le rôle des pages pivots est de référencer les sources d'information du Web. L'originalité de ce projet est la conception d'un modèle basé sur les relations entre ces deux types de pages. Pour l'équipe de ce projet, il existe un certain équilibre entre ces deux fonctions dans la structure du graphe du Web.

Un algorithme itératif nommé *HITS (Hypertext Induced Topic Search)* [Kleinberg, 1999] permet de calculer les valeurs de pivot et de référence de chaque page.
Sa phase initiale consiste à former pour chaque requête sur un sujet donné, un ensemble limité de pages qui a les propriétés suivantes : (i) l'ensemble doit être relativement petit , (ii) doit être riche en pages pertinentes et (iii) doit contenir de nombreuses pages références. Clever obtient une première liste de pages en utilisant un moteur classique comme Altavista. Pour les auteurs, cet ensemble vérifie le critère (i) et (ii) puisque l'ensemble des pages trouvées contient les termes de la requête, ce qui peut être discutable. Par contre le critère (iii) n'est pas forcement vérifié. Pour obtenir davantage de pages références, Clever ajoute aux pages trouvées, toutes les pages qui pointent vers elles ainsi que les pages qu'elles citent.
HITS évalue ensuite les fonctions pivot et référence de chaque page en utilisant le principe suivant : les bonnes pages pivots sont celles qui pointent vers de bonnes pages références et les bonnes pages références sont citées par de bons pivots. Des notes initiales sont affectées à chacune des pages pour ces deux fonctions. Elles sont affinées par un processus itératif en deux étapes :

- dans un premier temps l'algorithme utilise les notes de la fonction référence pour améliorer les notes de la fonction pivot ; les pages qui pointent vers les meilleures pages références voient leur note de pivot augmenter ;
- de la même manière, la fonction référence est réévaluée en utilisant les notes de pivots.

Ce processus converge rapidement. Les notes se stabilisent en cinq ou six itérations pour une ensemble de 300 pages environ.

Contrairement au Page Rank les notes pivot et référence sont dépendantes de la requête et réévaluées à chaque fois. En bibliométrie, on pourrait aussi distinguer deux types de documents : les documents qui font autorité sont les articles de recherche, tandis que les surveys (états de l'art) sont semblables aux

hubs car ils citent de nombreux articles de recherche.

2.3.3.3 La Catégorisation de pages

La catégorisation des pages vise l'organisation des corpus en affectant les pages dans des classes. Celle-ci peut se faire de manière supervisée ou non. Dans le premier cas, l'ensemble des classes est prédéfini et les méthodes sont dites de *classification*, alors que dans le deuxième cas il ne l'est pas et l'on parle de *clusterisation*.

En vue d'obtenir des répertoires du type *Yahoo*, de nombreux travaux se penchent sur la construction de classifieurs automatiques utilisant des méthodes apprentissages [Chakrabarti et al., 1998b], [Chakrabarti et al., 1998a]. Ces classifieurs se basent en priorité sur le contenu lexical des pages, l'analyse des liens permettant de compléter cette étude lexicale en analysant le voisinage des pages.[25]

Les différentes méthodes de clusterisation proviennent des statistiques multidimensionnelles et de l'étude des graphes. Les métriques utilisées combinent souvent l'approche lexicale et l'approche citationiste. Quelques auteurs se limitent uniquement à l'étude du graphe et transposent la méthode des co-citations sur le Web [Larson, 1996], [Prime et al., 2002b], [Pirolli et al., 1996]. Nous commenterons les résultats obtenus par ces méthodes au début du chapitre 4 avant de présenter notre méthode d'extraction de corpus homogènes basée sur le principe de co-citation.

2.3.3.4 La découverte de communautés d'intérêt

Un autre axe de l'analyse des liens vise la découverte de communautés d'intérêt présentes sur le Web. Kumar et al. évoquent trois motivations pour l'extraction de ces communautés [Kumar et al., 1999] : au niveau documentaire, ces communautés fournissent pour un domaine donné les ressources les plus fiables et les plus récentes ; au niveau sociologique, elles représentent les groupes sociaux présents sur la Toile, ce qui permet de suivre l'évolution de la structure intellectuelle du Web ; enfin, elles permettent de déterminer des cibles pour les campagnes publicitaires. Les différents auteurs s'intéressant à cette question définissent une cybercommunauté comme un ensemble de pages partageant le même sujet. Comme le souligne Bennouas et al. [Bennouas et al., 2003] une telle définition est difficilement utilisable car elle fait appel à la subjectivité

[25] le voisinage d'une page x est généralement un ensemble où chaque page peut atteindre ou être atteinte par la page x avec un nombre d'intermédiaires fixé.

de chacun. Si certains auteurs de pages web ont une volonté explicite de structurer la communauté (webrings[26], ect.), la plupart n'ont même pas conscience de leur communauté d'appartenance.

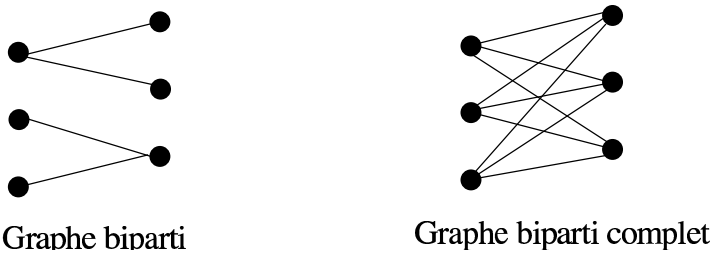

FIG. 2.14 – Exemples de graphes bipartis

D'un point de vue empirique, plusieurs auteurs définissent leur propre notion de cybercommunauté.

- Les approches suggérées par Gibson et al. [Gibson et al., 1998] et Kumar et al. [Kumar et al., 1999] reposent sur la découverte de sous-graphes bipartis. Un graphe est appelé *biparti* si l'ensemble de ses sommets se répartissent en deux sous-ensembles, et si toute arrête de ce graphe a une extrémité dans chacun de ces sous-ensembles (figure 2.14). Plus particulièrement, l'approche de Gibson et al. s'appuie sur l'algorithme HITS (section 2.3.3.2). Dans [Gibson et al., 1998], une communauté d'intérêt correspond aux n-premières pages références et aux n-premières pages pivots. Une des limites de l'algorithme HITS est le retour de pages concernant des thèmes voisins de la requête initiale. Gibson et al. propose une méthode pour la détection automatique de ces communautés associées.
  La méthode de Kumar et al. vise la détection de communautés émergentes. Ces communautés sont détectées par leur *signature*. Cette méthode repose sur l'hypothèse selon laquelle la signature d'une communauté peut être représentée par un graphe biparti avec une densité de liens importante.

- Dans [Flake et al., 2002], les cybercommunautés sont des collections de pages qui partagent davantage de relations entre elles qu'avec les autres pages présentes sur la Toile. On retrouve ici un rapprochement direct avec le concept de LSSets (section 2.1.4), groupe dans lequel les relations internes sont plus nombreuses que les relations externes. Les expérimentations menées par Flake et al. montrent que la découverte de communautés selon ce principe forme des ensembles de pages parta-

[26] Les webrings sont des ensembles de sites partageant le même thème et volontairement reliés les uns aux autres pour permettre la navigation d'un site à l'autre. Cette navigation s'effectue généralement au niveau des pages d'accueil qui contiennent un menu déroulant permettant de sélectionner le site désiré.

geant les mêmes thématiques.

– Bennouas et al. [Bennouas et al., 2003] proposent un modèle gravitationnel du Web pour la détection de communautés d'intérêt. Ce modèle très original s'inspire des lois physiques et en particulier des lois cosmologiques. Selon les auteurs, ces lois produisent de bonnes métaphores pour décrire le monde du Web. Dans cette approche, les pages sont modélisées comme des particules massives et les liens hypertextes comme des forces gravitationnelles. Ce modèle montre que les pages ont tendance à se regrouper en galaxies au gré des forces gravitationnelles subies. Ces galaxies contiennent des pages du même thème et sont spatialement proches les unes des autres : elles forment des cybercommunautés. Les auteurs proposent une autre analogie : l'autorité des pages est analogue à la masse des particules : une page de référence (faisant autorité) se comporte comme un soleil immobile autour d'un nuage de planètes, c'est-à-dire des pages ayant trait au même sujet.

Chapitre 3

Notre approche de caractérisation des documents

Ce chapitre présente notre approche de caractérisation des documents présents sur la Toile. Traditionnellement en documentation, les tâches de caractérisation de documents (indexation et catalogage) s'effectuent pour chaque document pris séparément. Dans cette thèse, nous envisageons une caractérisation collective des documents. Notre démarche vise l'organisation préalable des corpus : les documents proches (partageant des propriétés communes) sont regroupés en sous-ensembles. La caractérisation consiste à identifier et à renseigner les caractéristiques communes des documents de chaque sous-ensemble.

Notre appoche de caractérisation collective comporte deux étapes :
- l'extraction de corpus homogènes s'appuyant sur la structure hypertexte de la Toile,
- la qualification collective des documents basée sur leurs contenus.

Ce chapitre s'organise de la manière suivante. Dans un premier temps, nous précisons les notions de document web et de site web, puis nous présentons nos choix méthodologiques pour l'extraction de corpus homogènes et la qualification des documents.

3.1 Notions de documents, de sites

Lorsque l'on s'intéresse à la caractérisation des documents sur le Web, il est inévitable de s'interroger sur la notion même de **document web**. Très souvent, celui-ci est assimilé à une page web. C'est pourquoi, pour la plupart des systèmes de recherche d'information, les unités informationnelles retournées aux utilisateurs sont les pages web. Celles-ci, nœuds du réseau hypertexte sont

des composantes élémentaires généralement autonomes. Si d'un point de vue sémantique elles se suffisent à elles-même, elles ne correspondent pas toujours à la totalité d'un document au sens traditionnel. En effet, la vocation d'une page web est d'être lue en ligne et les recommandations concernant sa longueur maximum varient entre un ou deux écrans traditionnel de 17 pouces. L'hypertextualisation des documents traditionnels les plus longs est donc effectuée sur plusieurs pages. Un document traditionnel peut être : une partie d'une page web, une page web tout entière ou un regroupement de plusieurs pages.

De manière générale, la notion de document dans les hypertextes est une question en soi. Selon le principe de multiplicité et d'emboîtement[1] énoncé par P. Levy [Lévy, 1990], un document hypertexte peut contenir d'autres documents hypertextes. De plus, les nœuds du réseau représentent des unités documentaires de niveaux différents (principe d'hétérogénéité[2]) et ne sont pas comparables.

S'il est difficile et peut-être vain de vouloir définir un document web au sens traditionel, il existe sur le Web des ensembles de pages homogènes d'un point de vue documentaire et repérables facilement. Il s'agit des **sites web**, ensembles cohérents de pages (objectifs et thèmes communs), créés et maintenus par une même autorité. Au niveau de la forme, les pages d'un site partagent la même charte graphique et ceux-ci possèdent toujours une page d'accueil, point d'entrée permettant d'atteindre les ressources du site. Les sites web sont des systèmes hypertextes autonomes. Ils sont comparables aux magazines car ils proposent des ressources hétérogènes. Le Web est un donc un système englobant une multitude de systèmes hypertextes (les sites web), eux-mêmes reliés les uns aux autres par des liens hypertextes.

3.2 Extraction de corpus homogènes

Il existe plusieurs méthodes issues des statistiques et de l'analyse de données permettant de structurer les corpus (les analyses factorielles, les MDS (*Multi-Dimensional Scaling*), les classifications). Comme en bibliométrie, deux orientations sont possibles sur le Web pour le choix des métriques sous-jacentes à ces méthodes de structuration : une orientation lexicale, utilisant les mots contenus sur les pages et une orientation citationiste, basée sur les relations

[1] Le principe de multiplicité et d'emboîtement est un des six principes qui selon Pierre Lévy caractérisent les réseaux hypertextuels. Pour cet auteur, l'ensemble hypertextuel est un réseau d'associations qui permet des encastrements de documents. Par exemple, une critique de livre peut contenir un lien vers la biographie de l'auteur. Ainsi cette biographie est emboitée dans la critique de livre.

[2] Le principe d'hétérogénéité recouvre différents aspects : hétérogénéité des nœuds (texte, image, son ou vidéo), hétérogénéité des liens, des unités documentaires représentées par les nœuds (par un exemple un sommaire, un chapitre, une section ou un paragraphe), etc.

entre les pages matérialisées par des liens hypertextes. Dans notre approche, l'organisation des corpus utilise une métrique basée sur les liens. Elle repose sur l'hypothèse d'une auto-organisation de la Toile analogue à l'auto-organisation de l'univers des publications scientifiques.

3.2.1 L'hypothèse d'une auto-organisation de la Toile

L'univers des publications scientifiques et le Web sont des espaces multi-auteurs dans lesquels les documents peuvent être reliés les uns aux autres. Ceci implique une structure, une auto-organisation de ces univers et permet différents parcours dans les corpus. Les raisons qui amènent un scientifique à mentionner les travaux de ses prédécesseurs sont discutées dans la section 2.2.4.2. Rapellons toutefois que la citation est un acte socio-cognitif relativement normé et contrôlé : il arrive qu'un article soumis à une revue soit rejeté simplement pour l'oubli d'une référence à un article pertinent.
De son côté, la création de liens hypertextes est moins formelle et n'est soumise à aucun contrôle. Soucieux de mener à bien l'analogie entre l'analyse des citations et le Web, différents scientomètres [Thelwall, 2003], [Chu, 2004], [Kim, 2000] s'interrogent sur les raisons qui poussent un auteur à créer un lien hypertexte. Les études menées par Thelwall [Thelwall, 2003] et Chu [Chu, 2004] font la différence entre :

- les liens intra-sites : liens reliant des pages hébergées sur le même site,
- et les liens inter-sites : liens reliant des pages hébergées sur des sites différents.

Les liens internes aux sites relient des pages créées par une même autorité. Ils servent d'une part, à la structuration des documents (liens entre les différentes parties d'un document), et d'autre part, à la navigation ou à la citation entre les ressources d'un même site. La citation entre les pages d'un même site s'apparente à l'auto-citation et n'indique pas nécessairement de « liaison remarquable » entre les pages citantes et citées. Un lien vers une page hébergée sur un autre site, invite l'utilisateur à quitter le site sur lequel il se trouve pour en visiter un autre. Ce genre d'invitation est généralement légitimée par l'intérêt que l'auteur porte à l'autre site. Le terme *sitation*[3] est alors inventé pour désigner ce type de liens.

Les résultats obtenus par Thelwall [Thelwall, 2003] montrent que les motivations de *sitation* sont assez différentes des motivations de citation. Son étude porte sur les raisons de *sitation* des universités anglaises. Le corpus utilisé comporte 111 sites d'université et 19.438 liens pointant vers les pages d'accueil de ces universités. Parmi les 19.438 liens, 100 sont tirés au hasard pour cette expérimentation. Une analyse qualitative tente d'expliquer quelles sont les motivations

[3]Le terme *sitation* est employé dès 1997 par Rousseau [Rousseau, 1997] ; il s'agit de la contraction de l'expression anglaise « site citation ».

à l'origine de ces liens. A l'issue de ces observations, l'auteur définit quatre types de liens possibles.

1. Les liens de navigation générale. Ils permettent la navigation vers des informations que l'auteur considère intéressantes, hébergées sur d'autres sites. Ces liens sont qualifiés de *navigation générale*, car les informations retrouvées sur la page cible ne partagent pas le thème de la page source.
2. Les liens de propriété. Ils permettent de revendiquer la propriété intellectuelle d'un document. Dans cette étude, ce type de lien est retrouvé sur les pages de projets collaboratifs entre plusieurs établissements. En effet, les documents relatifs à ce type de projet sont généralement regroupés en un seul endroit, souvent sur le site web d'un des participants. C'est pourquoi, il est important de mentionner sur les pages web de ces projets l'ensemble des participants.
3. Les liens sociaux. Ce sont des liens vers des partenaires, des collaborateurs.
4. Les liens gratuits. Liens, qui selon Thelwall, sont sans motivation de communication particulière. Par exemple, des liens vers l'université où l'on a suivi ses études, des liens vers une ancienne entreprise, etc.

Nous pensons qu'une telle démarche ne peut donner une taxinomie des motivations de *sitation* exhaustive. En effet, les pages du corpus (vers lesquelles les liens pointent) sont d'un genre particulier. Il s'agit de pages d'accueil d'université. En général, ces pages sont analogues aux sommaires : elles donnent l'accès aux ressources du site et présentent rarement des informations de fond. Ceci explique pourquoi aucun lien cognitif, ou du moins aucun lien de navigation thématique, n'apparaît dans cette expérience. Sans prétendre l'exhaustivité, nous proposons de compléter cette classification avec :

- les liens de navigation thématique, permettant la navigation entre pages du même thème,
- et les liens cognitifs, qui pointent vers des pages évoquant ou argumentant les idées de la page initiale.

De plus, parmi les *liens gratuits*, c'est-à-dire ceux qui apparaissent sans motivation de communication particulière (entre les pages citantes et citées), nous pourrions mentionner les liens de publicité : *liens gratuits* dans le sens où ils n'apportent rien d'un point vue social ou sémantique, mais qui rapportent financièrement.

L'étude de H. Chu [Chu, 2004] confirme aussi que les raisons de *sitation* sont assez différentes et surtout moins complexes que les motivations de citation. La *sitation* sert généralement à mentionner un site intéressant. Elle est rarement utilisée pour argumenter, comparer ou présenter des idées. D'autre part, elle est moins précise. Alors qu'un article peut être cité pour l'intérêt porté à un seul paragraphe ou une phrase bien précise, la *sitation* vise généralement au moins une page et souvent le contenu d'un site tout entier.

Au niveau de la recherche d'information, les liens particulièrement intéressants sont :

- les liens de navigation générale, car ils indiquent un chemin possible pour accéder aux ressources,
- les liens de navigation thématique et les liens cognitifs, puisqu'ils mettent en relations des pages de même centre d'intérêt.

Les liens sociaux peuvent jouer un rôle important pour une compréhension plus globale de la Toile. Malheureusement, nul ne peut prétendre à un typage systématique des liens. D'une part, car il s'agit d'une tâche extrêmement complexe qui ne peut s'envisager automatiquement, et d'autre part, compte tenu du caractère multi-auteurs du Web, de la diversité de leurs besoins de communication, les liens émis traduisent des motivations *ad-hoc* selon l'information manipulée [Aguiar, 2002].

3.2.2 Relations intéressantes dans le graphe du Web

Dans cette section, nous allons examiner différentes relations susceptibles de lier des pages partageant des propriétés communes comme le thème, la portée géographique, la cible, le niveau, le genre/type ou la langue. Les trois relations étudiées sont la relation de *sitation*, la relation de couplage et la relation de *co-sitation* (figure 3.1).

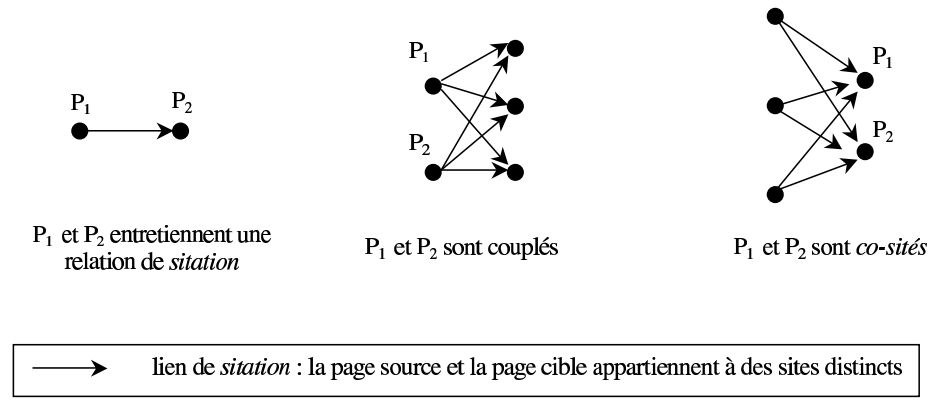

FIG. 3.1 – Trois relations possibles entre les pages web

1. La relation de *sitation*
   Nous formulons l'hypothèse H_1 suivante.

   > H_1 **:** Si une page P_1 contient un lien hypertexte vers une page P_2 (avec P_1 et P_2 hébergées sur deux sites distincts), il existe pour le créateur de la page P_1 une motivation pour citer la page P_2 à partir de cette page P_1 ; et cette motivation se traduit par des propriétés communes entre les deux pages.

Nous avons constaté dans la section précédente que les motivations de *sitation* sont très différentes d'un lien à l'autre. Alors que l'hypothèse H_1 semble assez bien vérifiée pour les liens thématiques ou les liens cognitifs, sa validité est moins évidente pour les liens de navigation générale. En effet, deux pages reliées par un lien cognitif ou thématique, partagent le même thème, s'adressent généralement aux même interlocuteurs, sont probablement du même type (genre) ou écrites dans la même langue. Même si elles peuvent avoir une langue commune ou la même portée géographique, deux pages reliés par un lien de navigation ne partagent pas le même thème et sont rarement du même type. En effet, les liens de navigation sont souvent issus de pages organisant les ressources du Web (comme les répertoires (ou annuaires)), alors que les pages recevant ce type de lien sont plutôt des pages de contenu (contenant des informations de fond).

2. La relation de couplage
   Par analogie avec la méthode du couplage bibliographique provenant de la bibliométrie (section 2.2.3.2), deux pages web entretiennent une relation de couplage (sont couplées) si elles partagent des liens hypertextes identiques. Les pages fortement couplées (partageant de nombreux liens identiques) ont toutes les chances de détenir des propriétés communes, et pour cause, ce sont généralement des pages doublons[4]. De manière générale, le couplage favorise le rapprochement de pages émettant de nombreux liens (les répertoires, par exemple). Les pages n'ayant aucun lien externe, comme c'est souvent le cas pour les pages de contenu, ne sont jamais couplées avec d'autres. De telles pages ne pourraient pas être classées par une méthode de structuration utilisant une métrique basée sur la relation de couplage. Ce qui interdit l'usage du couplage dans le cadre d'une application de recherche d'information.

3. La relation de *co-sitation*
   La relation de *co-sitation* est la relation existante entre deux pages citées simultanément par une page hébergée sur un site différent des pages citées. Nous formulons l'hypothèse H_2 suivante.

   > H_2 **:** Si une page P contient un lien hypertexte vers les pages P_1 et P_2, il existe (au moins pour le créateur de la page P) une raison pour citer ces deux pages ensemble. L'association existante entre les deux pages P_1 et P_2 est d'autant plus forte, si elle est reprise par d'autres auteurs et si les pages P_1 et P_2 sont souvent citées ensemble par rapport à leurs fréquences de citation respectives. Cette association se traduit par des propriétés communes entre les pages.

   L'hypothèse H_2 nous parait valide pour de nombreux cas. En effet, comme l'équipe du projet Clever ([Gibson et al., 1998], [Kleinberg, 1999]), nous

[4] Différentes méthodes ([Bharat et al., 2000], [Prime et al., 2002a]) permettant la détection de sites miroirs sont basées sur la force de couplage (nombre de liens sortant communs).

pensons qu'il existe une auto-organisation du Web avec un équilibre entre des pages *pivots* et des pages *références*. Les pages pivots sont celles qui organisent la navigation sur le Web. Elles émettent des liens de navigation générale vers des ressources. Généralement, ces pages sont formées intelligemment et pointent vers des pages références ayant entre elles des points communs. D'autre part, les liens de *sitation* présents sur les pages références (généralement des pages contenant de l'information de fond) peuvent être de même nature : liens de navigation thématique, liens cognitifs, liens sociaux. Par l'intermédiaire de la *co-sitation*, ces liens peuvent lier des pages aux caractéristiques communes.

Des trois relations étudiées, la *co-sitation* semble la plus apte à rapprocher des pages aux caractéristiques communes. Pour l'extraction de corpus homogènes, nous orientons notre choix vers l'utilisation d'une métrique basée sur la relation de *co-sitation*.

3.2.3 Limites

Dans cette section, nous présentons les limites de l'analyse des *sitations* et plus particulièrement du principe de *co-sitation*.

- Un des problèmes majeurs concerne les *sitations* vides de sens (les liens gratuits recensés par Thelwall, les liens de publicité). Alors qu'en bibliométrie classique les éléments les plus cités sont structurants (contrairement aux mots), qu'en est-il sur le Web où de relations de publicité sont très présentes ? Retrouve-t-on la même configuration que pour les mots où les éléments les plus fréquents font partie du bruit ? Si des études confirmaient cette hypothèse, une sélection bradfordienne [Bradford, 1934] pourrait être envisagée pour éliminer les éléments les plus cités (le bruit). Cependant, les liens gratuits apparaissant avec une fréquence faible ne pourraient être supprimés.

- Au sein des sites web, toutes les pages sont généralement citées au moins une fois. Ceci permet la navigation complète dans le site et la lecture de toutes les pages. Par contre, les pages recevant des citations externes (pages recevant des *sitations*) sont nettement moins nombreuses. En effet, chaque site comporte seulement quelques points d'entrée. Les points d'entrée sont des pages par lesquelles l'arrivée sur le site se fait de manière cohérente. Ces pages se retrouvent à différents niveaux des sites : au sommet bien sûr avec les pages d'accueil, mais aussi aux niveaux inférieurs avec les pages représentant le début d'un document logique. Les résultats de l'étude de Chu [Chu, 2004] indiquent que la *sitation* n'est pas très précise et qu'elle vise souvent le contenu d'un site tout entier. Ainsi ce sont souvent les pages d'accueil qui sont citées. Toutefois, la *sitation* de pages autres que des points d'entrée est tout à fait possible, mais rare. C'est pourquoi, l'analyse des *sitations* s'intéresse principalement aux liens reçus par les points d'entrée de chaque site.

Les méthodes basées sur la *sitation* (i.e sur les liens externes) appliquées au Web, prennent surtout en compte les points d'entrée et non la totalité des pages web.

– Une autre limite concerne la transposition du principe de co-citation sur le Web. En bibliométrie classique, les distributions des degrés sortants (citations émises) suivent des lois gaussiennes. Le nombre de références bibliographiques par article variant entre 20 et 50 (section 2.2.2.2, page 29). Sur le Web, les distributions des degrés sortants suivent des lois hyperboliques. La probabilité pour qu'une page n'émette qu'une seule *sitation* est relativement forte. Ainsi, il est tout à fait envisageable qu'un point d'entrée, même très cité, ne soit jamais co-cité. Autrement dit, il est possible qu'une page soit toujours citée par des pages n'émettant qu'une seule *sitation*.

La mise en œuvre des méthodes doit surmonter différents problèmes.

– L'analyse des *sitations* est basée sur les relations inter-sites. Pourtant, il est bien difficile d'identifier automatiquement les sites. Deux approximations possibles consistent à dire qu'un site correspond
  * aux pages hébergées sous un nom de domaine (par exemple `emse.fr`),
  * ou aux pages hébergées sous la concaténation d'un nom de machine et d'un nom de domaine (par exemple `www.emse.fr` ou `rim.emse.fr`).

  Ces deux approximations engendrent des erreurs. D'une part, plusieurs sites peuvent coexister sous le même nom de domaine. La plupart des fournisseurs d'accès comme Wanadoo[5] ou Free[6] proposent d'héberger sous leur nom de domaine les sites de leurs clients[7]. D'autre part, les sites importants peuvent être hébergés sur plusieurs machines.
– Une autre difficulté réside dans l'obtention d'un graphe du Web propre. La découverte du graphe se fait en parcourant la Toile de liens en liens. Il faut pouvoir gérer :
  * les erreurs présentes dans l'utilisation de la syntaxe HTML et dans l'écriture des URLs. Il faut choisir si l'on tient compte des pages et des URLs écrites de manière imparfaite. Voici quelques exemples d'erreurs possibles :
    1. <A<A HREF="http ://www.xyz.fr/">xyz</A>
    2. `htp://www.xyz.com.fr/`
    3. `www.xyz.com.fr/`

    1) présence de deux chevrons ouvrants au lieu d'un seul ; 2) erreur de frappe dans le protocole ; 3) oubli du protocole, ce qui est compris comme un lien relatif.
  * les URLs alias, c'est-à-dire les différents liens possibles pour désigner un même fichier source. L'existence des URLs alias provient à la fois

[5] `http://www.wanadoo.fr/`

[6] `http://www.free.fr/`

[7] Les noms de machines sont parfois différents pour chaque site comme c'est le cas pour Free (par exemple `http://aquafish.free.fr/`), ou au contraire tous les sites personnels sont regroupés sous une même machine comme pour Wanadoo (`http://perso.wanadoo.fr/`)

des noms de domaines alias (par exemple `av.com` est un alias du nom de domaine `altavisa.com`), et des liens alias (raccourcis) créés sur les serveurs hébergeant les sites.

* les sites dupliqués (i.e. les sites miroirs).

Finalement, il n'existe pas un graphe du Web, mais plusieurs graphes possibles en fonction des choix méthodologiques que l'on fait pour la découverte du graphe.

3.3 Qualification des pages et des sites web

Dans la section précédente, nous avons discuté des possibilités et des limites d'une structuration basée sur l'analyse des liens. Cette section présente nos choix pour la qualification des pages et des sites web.

3.3.1 Choix des métadonnées

Parmi les métadonnées proposées par le Dublin Core (section 1.3.2, page 9), nous pensons que ce sont les métadonnées Sujet, Type, Couverture spatio-temporelle et Langage qui seraient utiles pour améliorer la recherche d'information sur le Web. En effet, vu l'abondance des ressources disponibles et de leurs diversités, l'utilisateur n'a pas a priori d'informations précises sur les auteurs des ressources qu'il recherche, ni sur leurs dates d'édition. Par contre, il connaît le thème, le type, la portée spatio-temporelle des documents recherchés, et les langues qu'il est capable de lire.

Alors que les discussions menées par la communauté informatique autour des standards de métadonnées sont très riches, celles concernant leur affectation et les difficultés liées à cette tâche sont quasi-inexistantes. Bien peu proposent des normes ou des listes de contrôle pour les valuer. Concernant le champ thématique des documents, on peut facilement se rapprocher des travaux anciens en sciences de l'information sur les thesauri et les langages à facettes, ou encore du domaine de l'informatique avec les ontologies. De plus, nombreuses sont les méthodes qui tentent de découvrir automatiquement le thème des documents [Chakrabarti et al., 1998b], [Chakrabarti et al., 1998a], [Chekuri et al., 1996]. Par ailleurs, différents outils automatiques permettent aussi de détecter rapidement la langue d'un document. Ding et al. quant à eux, proposent une méthode pour déterminer la portée géographique des pages [Ding et al., 2000]. Finalement, le type des documents reste selon nous une des propriétés particulièrement intéressante à connaître dans le cadre d'une recherche documentaire, et pourtant très peu investie.

Quelques chercheurs se penchent sur le genre ou le type de documents que l'on retrouve sur le Web. Plus particulièrement, Crowston et Williams [Crowston and Williams, 2000] s'intéressent aux différents genres de documents reproduits ou émergents sur la Toile. Cette étude, qui débute en 1996, porte sur un corpus de 1.000 URLs tirées au hasard parmi une liste de 8.000 URLs. Cette liste, fournie par l'équipe du moteur de recherche Altavista[8], est elle-même formée de manière aléatoire. Les pages correspondant aux URLs (fichiers HTML et images) sont rapatriées grâce à un robot[9]. Un codage est effectué par un assistant pour déterminer le genre de chaque page. Pour valider ce codage, 40% des pages approximativement, sont elles-aussi examinées par un des auteurs. Ce double codage aboutit à des résultats identiques pour 68% des pages, pour 10% des pages les codages sont similaires (proches) et pour les autres pages deux codages sont possibles. Cette étude confirme qu'une page ne correspond pas toujours à un document. Dans la plupart des cas, ce n'est pas le genre de la page qui est identifié, mais celui du document dont elle fait partie. L'étude montre que dans 60% des cas, les pages appartiennent à un genre familier. Soit le genre est complètement reproduit, comme c'est le cas pour les articles, les FAQ ou les programmes de cours. Soit le genre est connu, mais adapté au nouveau support que constitue le Web. Les genres adaptés utilisent pleinement les différentes possibilités offertes par l'hypermédia. Par exemple, les arbres généalogiques ne sont pas reproduits de manière hiérarchique, mais éclatés sur plusieurs pages, des liens permettant la navigation entre les générations. Ce qui nous intéresse tout particulièrement dans cette section est l'identification des genres émergents (environ 30% des pages). Ces nouveaux genres sont :

- les *homepages* : pages concernant une personne ou une organisation et la présentant,
- les listes et les pages thématiques : pages comportant de nombreux liens et organisant des ressources du même thème,
- les pages des serveurs web : pages donnant des informations relatives au serveur sur lequel elle est hébergée (comme des statistiques sur la consultation du serveur),
- les pages interactives, permettant l'échange de données avec des utilisateurs (comme les formulaires d'interrogation des moteurs de recherche).

Notons que dans 10% des cas, les pages appartiennent à des genres non identifiés ou pas encore reconnus.

Pirolli et al. [Pirolli et al., 1996] s'intéressent eux-aussi aux types de pages présents sur la Toile. Dans leur article, les auteurs présentent à la fois une typologie des pages web et un dispositif permettant d'affecter un type à chacune des pages. La typologie suggérée par les auteurs est la suivante[10].

- *Head* : correspond aux points d'entrée tels que nous les avons définis dans la section 3.2.3. Ce sont des pages qu'il est préférable de consulter en premier car elles suggèrent une navigation vers d'autres pages. Selon

[8] `http://www.altavista.com`

[9] Sur les 1.000 URLs, 837 pages sont retrouvées, 128 URLs sont obsolètes (erreur 404) et 35 appartiennent à des sites où le temps de réponse est trop long.

[10] Certains termes de la typologie (en italique) sont volontairement non traduits, par peur d'imprécision.

les auteurs, ce type de page se divise en deux sous-classes :
  * *homepages* organisationnelles : point d'entrée sur les sites des organisations,
  * *homepages* personnelles : pages personnelles des individus,
- *Index* : pages comportant de nombreux liens et permettant la navigation vers d'autres pages,
- *Source index* : pages index dans des espaces relatifs et correspondant aussi à des points d'entrée,
- Référence : pages utilisées à plusieurs reprises pour expliquer un concept ou qui correspondent à une référence effective. Les pages références se partagent en deux sous-classes :
  * destination : pages très citées mais qui ne pointent nulle part (références bibliographiques, copyright),
  * contenu : pages dont l'objectif est de délivrer de l'information.

Notons que l'on retrouve dans cette typologie certains des genres émergents proposés par Crowston et Williams, comme les homepages, les listes (nommées index par Pirroli et al.). La typologie de Pirroli et al. s'appuie plus sur le rôle que jouent les pages dans le graphe, que sur le type d'information que l'on y trouve. D'ailleurs, les données du graphe sont un des éléments exploités par le dispositif affectant un type aux pages. Ce dispositif utilise tous les types d'information liés au Web et à son usage comme : le *contenu* des pages, c'est-à-dire le texte et les méta-informations (taille des fichiers, par exemple), le *graphe*, les données provenant de l'*usage*. Il se base sur un ensemble de règles déterminant quel type doit être attribué aux pages. Voici quelques exemples de ces règles : une page ayant de nombreux liens sortants et une faible taille par rapport à son nombre de liens est un *index* ; une page de taille importante ayant peu de liens entrants et sortants est une page de contenu ; etc.

3.3.2 Proposition d'une typologie des sites et pages web

Contrairement à Crowston et Williams, notre objectif n'est pas la détermination du genre des documents logiques (i.e. des articles, des FAQ, des thèses, etc.) disponibles sur la Toile, car, d'une part, il est bien difficile de déterminer et de repérer ces documents ; d'autre part, les documents logiques (en particulier les documents reproduits) sont de plus en plus souvent disponibles sous différents formats[11] plus adaptés à la lecture et à l'impression que le format HTML, et se situent donc à la frontière du Web.

La typologie que nous proposons dans cette section est relative aux sites et aux pages web. Nous définissons quatre métadonnées qui sont : le *Type d'autorité* responsable du site, le *Type de site*, le *Type de page* et le *Type d'infor-*

[11] Quelque exemples de ces formats sont : le format *Postscript* (PS), le *Portable Document Format* (PDF) ou le *Rich Text Format* (RTF).

mation contenue sur la page. Cette typologie est une approche personnelle qui s'inspire des résultats des travaux antérieurs ([Crowston and Williams, 2000], [Pirolli et al., 1996]) et qui, bien sûr, peut être amenée à évoluer. Les termes en gras représentent les différentes valeurs possibles pour les métadonnées.

3.3.2.1 Type d'autorité

Pour mieux comprendre l'apport informationnel d'un site, savoir qui est à l'initiative de sa création peut être un indice important. Nous avons distingué quatre grandes classes possibles pour cette métadonnée qui sont :

- l'**institution**, c'est-à-dire un organisme généralement public, établi dans une société donnée pour répondre à un besoin particulier, qui revêt une valeur officielle ou légale[12],
- l'**entreprise**,
- l'**association**, au sens « groupement de personnes réunies dans un dessein commun non lucratif » (définition issue du dictionnaire Larousse),
- la **personne individuelle** et plus largement le groupement de personnes non officiel (membres d'une même famille, des amis, etc.).

Des sous-classes peuvent être définies en fonction du corpus à analyser. Par exemple, pour l'institution nous pouvons distinguer les établissements scolaires, les musées, les ministères, etc.

3.3.2.2 Type de site

Le Type de site dépend du rôle informationnel que veut jouer le site. Nous avons recensé 4 types distincts.

- Le plus courant, le site vitrine (**homeserveur**) favorise l'information autodescriptive, celle décrivant l'autorité responsable du site. Sorte de « plaquette », l'objectif premier pour les auteurs de ces sites est de se présenter. Les thèmes abordés en priorité sont : Qui sommes-nous ? Nos activités, nos produits, nos partenaires, comment nous joindre, etc. Cependant, ces sites peuvent aussi héberger dans des niveaux inférieurs (à plusieurs « clics » de la page d'accueil) des documents non autodescriptifs.
- Le **site de recherche** propose un accès aux ressources du Web. Nous distinguons deux sous-classes pour ces sites de recherche : les moteurs de recherche et les annuaires. Les moteurs sont des outils qui indexent automatiquement les pages web, alors que les annuaires font intervenir des indexeurs humains qui généralement classent les points d'entrée en sections thématiques. Il existe des sites de recherche généraux et des sites de recherche spécialisés pour un ou plusieurs types d'information

[12] D'après l'office québécois de la langue française, `http://www.granddictionnaire.com/`.

particuliers. Des exemples de moteurs généraux sont Google[13] et Altavista[14] ; des exemples d'annuaires généraux sont Yahoo[15] et Lycos[16]. Le moteur Citeseer [17] n'indexant qu'un seul type de documents (les articles scientifiques) est un exemple de site de recherche spécialisé.
- Se comportant comme un éditeur, le **site de ressources** organise et propose ses ressources propres (contrairement aux sites de recherche). Ils se présentent souvent comme des bibliothèques ou des bases de données.
- Les **services web** proposent des services liés à la vie sur le Web et l'Internet, comme des messageries, forums de news, de l'IRC[18], etc.

3.3.2.3 Type de page

Les valeurs proposées pour le type de page sont assez proches de celles proposées par Pirolli et al. Elles dépendent surtout des caractéristiques physiques des pages. Nous avons retenu cinq valeurs possibles :
- la **page d'accueil**, point d'entrée principal (au sommet) du site,
- les **portails**, c'est-à-dire des pages comportant de nombreux liens sortants externes (vers d'autres sites),
- les **index**, organisant les ressources internes d'un site et comportant de nombreux liens sortants internes,
- les pages de **contenu**, pages comportant davantage de texte que de liens,
- les pages de **formulaire**, pages permettant d'interagir avec les utilisateurs.

3.3.2.4 Type d'information (contenue dans la page)

Sur les sites *homeserveur*, l'information prédominante concerne l'autorité du site et ses activités, alors que sur les autres types de sites (les sites de recherche, les sites de ressources et les services web) l'information autodescriptive est plutôt rare. Toutefois, l'information **autodescriptive** et **non autodescriptive** peuvent toutes les deux coexister au sein des sites. Connaître le type d'information contenue sur la page peut contribuer à améliorer la recherche d'information. Considérons le cas d'un utilisateur recherchant de l'information précise sur une organisation, un centre de recherche par exemple. Celui-ci consultera en priorité les pages du site officiel de cette organisation contenant de l'information autodescriptive. Par contre, s'il recherche plutôt des articles scientiques, il s'orientera vers les pages contenant de l'information non autodescriptive.

[13] `http://www.google.com/`
[14] `http://www.altavista.com/`
[15] `http://fr.yahoo.com/`
[16] `http://www.recherche.lycos.fr/annuaire/`
[17] `http://citeseer.ist.psu.edu/cs/`
[18] Internet Relay Chat.

3.4 Conclusion

Dans ce chapitre nous avons présenté nos choix méthodologiques pour l'extraction de corpus homogènes et la qualification des documents. Dans le chapitre quatre, nous reviendrons sur l'hypothèse d'une auto-organisation de la Toile. Nous étudierons la validité de l'hypothèse H_2 (section 3.2.2, page 60) pour les métadonnées présentées dans notre typologie des sites et des pages web. Le chapitre cinq, présente deux méthodes permettant la caractérisation collective des pages web. Ces méthodes mêlent l'affectation manuelle de métadonnées (utilisant le contenu des pages et des sites web), et l'affectation automatique par propagation.

Chapitre 4

Extraction de corpus ho mogènes par le principe de *co-sitation*

4.1 Objectifs du chapitre

Après avoir expliqué nos choix méthodologiques, notamment l'intérêt que nous portons à l'étude du graphe web par la méthode des *co-sitations*, l'objectif de ce nouveau chapitre est de montrer comment la structuration utilisant le principe de *co-sitation* permet d'organiser les corpus web et de regrouper des pages afin d'obtenir des sous-ensembles homogènes.

Différentes études menées sur le Web utilisant la méthode des co-citations (inter-sites ou intra-site) ont déjà montré qu'il était possible de rapprocher des pages ou sites web d'un même thème [Larson, 1996], [Prime et al., 2002b], ou au sein d'un même site des documents du même type [Pitkow and Pirolli, 1997].

Larson [Larson, 1996] est le premier à expérimenter la transposition de la méthode des co-citations sur le Web. Son objectif est de découvrir la structure intellectuelle du Web, c'est-à-dire de repérer automatiquement des domaines et des sous-domaines, sans avoir recours à l'indexation automatique. Son étude porte sur un corpus composé de 34 sites web concernant la géographie et les sciences de la terre (*geography, GIS, earth sciences*) sur lequel il transpose la méthode des co-citations d'auteurs [White and Griffith, 1981]. Il résout les difficultés dues à l'hétérogénéité du web en sélectionnant manuellement les sites pertinents et homogènes. Les citations faites vers des pages en dehors du sujet défini sont également supprimées manuellement. Les cartes thématiques obtenues par les techniques des MDS (*Multidimensional Scaling*) donnent des résultats clairs, raisonnables et interprétables.

Dans une étude similaire [Prime et al., 2002b], nous nous sommes intéressés à la transposition de la méthode des co-citations de documents (et non d'auteurs). Notre étude a porté sur un corpus de 3.500 pages concernant la bibliométrie, la scientométrie et les indicateurs scientifiques dans lequel aucune sélection manuelle n'a été effectuée. Elle met en évidence les limites théoriques et techniques de l'analogie, mais montre également l'intérêt de la structuration pour identifier les sous-thèmes de ce domaine et des domaines connexes.

L'expérience menée par Pitkow et Pirolli [Pitkow and Pirolli, 1997] se distingue par le choix du corpus utilisé : un unique et très grand site web, celui du « Georgia Institute of Technology's Graphic Visualization and Usability Center ». Les auteurs utilisent la méthode des co-citations de documents, et montrent comment celle-ci regroupe les pages non pas en fonction de leur thème mais en fonction de leur type, comme les projets de recherche (*Research Projects*), les pages personnelles (*People pages*), les documents de contenu (*Documents contents*), etc.

Ce chapitre décrit une expérience que nous avons menée au cours du deuxième semestre 2001. Notre objectif initial était d'étudier la structuration par le principe de *co-sitation* de documents, pour des corpus composés des pages issues de différents sites Web. Nous voulions montrer que celle-ci est capable de rapprocher non seulement des pages proches thématiquement, mais aussi des pages partageant d'autres propriétés : pages écrites dans la même langue, du même type, de la même origine géographique, etc. Finalement, dans le cadre de cette étude pour les raisons évoquées au chapitre précédent, nous nous sommes intéressés comme Pitkow et Pirolli aux propriétés typologiques des documents.

L'expérience présentée ici n'est pas un nouvel essai de transposition de la méthode des co-citations. En effet, même si notre méthode s'appuie sur le principe de *co-sitation*, l'objectif de notre expérience et la manière dont nous avons formé notre corpus diffèrent de la méthode classique.

4.2 Description du protocole

Cette expérience s'est déroulée en quatre étapes (fig. 4.1).

- La première étape est la constitution de notre corpus, base de test pour cette étude. Notre objectif est d'extraire un ensemble de pages partageant des propriétés typologiques différentes et dans lequel il existe de nombreuses relations de *co-sitation*.

Les deux étapes suivantes ont été menées en parallèle et indépendamment l'une de l'autre :

- d'un coté, nous avons effectué la structuration du corpus par la technique des *co-sitations* de pages,

– de l'autre, nous avons indexé (ou caractérisé, qualifié) manuellement notre corpus pour les quatre métadonnées (définies dans la section 3.3.2 du chapitre précédent) relatives au type de document.

Notre objectif est de montrer que la structuration basée sur les hyperliens permet d'obtenir des groupes de pages (*agrégats* ou *clusters* en anglais) dans lesquels il existe au moins une propriété typologique commune, sorte de point commun que partagent toutes les pages d'un groupe.

– Notre quatrième étape est donc une analyse quantitative de chaque agrégat pour mesurer l'homogénéité (la ressemblance) des pages qui le composent.

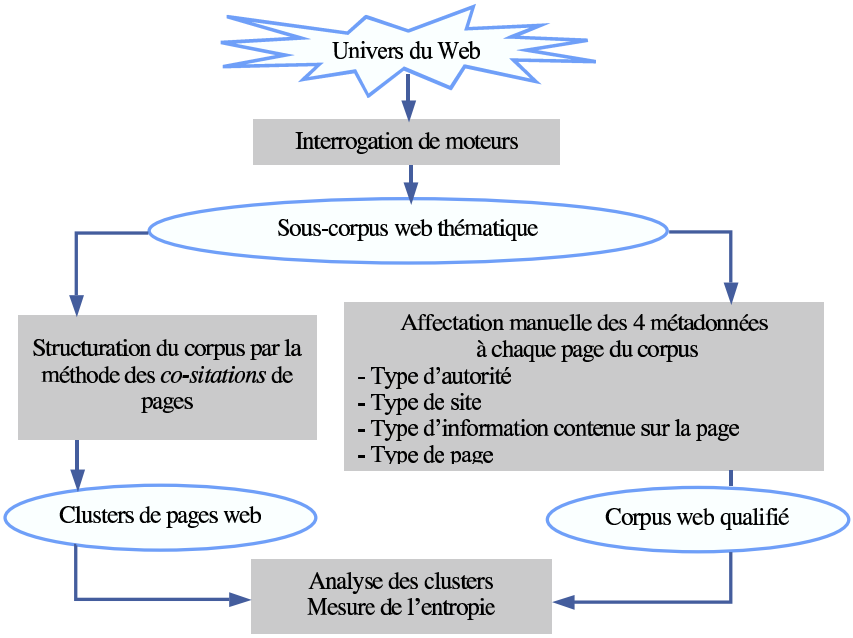

FIG. 4.1 – Etapes du processus expérimental

4.3 Constitution du corpus de test

4.3.1 Présentation de la méthode utilisée

Pour assurer un nombre important de relations de *co-sitation* entre les pages (une forte densité dans le graphe de *co-sitation*), nous avons choisi de former un corpus « mono-thématique » regroupant des pages écrites dans une même langue. Pour cette expérience, le thème retenu est l'astronomie, thème fédérateur engendrant de nombreuses publications de genre différent et provenant de communautés très diverses. Ceci nous garantit une bonne hétérogénéité pour les métadonnées présentées dans la typologie du chapitre trois. D'autre part, pour faciliter l'indexation, la langue choisie est le français.

Traditionnellement en bibliométrie, la construction des corpus s'effectue en interrogeant les bases de données de l'ISI[1]. La première étape consiste à rassembler les articles récents d'un domaine de recherche (les éléments citants, cf. paragraphe 2 de la section 2.2.3.2, page 36), la seconde à repérer quels sont les articles cités par ceux-ci (les éléments cités). Pour les bibliomètres, il s'agit de tâches lourdes et fastidieuses, qui passent par l'écriture de requêtes, l'interrogation des bases puis l'élimination du bruit et des doublons. Les corpus doivent être « propres » et exhaustifs ; la signification des cartes relationnelles en dépend.

Ces deux étapes permettent de construire la matrice de citation $M(n, m)$ qui donne les relations de citation entre les n articles citants et les m articles cités. La matrice de co-citation C s'obtient ensuite de la manière suivante :

$$C(m, m) = M^t(n, m) \times M(n, m). \tag{4.1}$$

La matrice de co-citation $C(m, m)$ donne pour chaque couple d'articles cités (i, j), sa fréquence de co-occurrence, c'est-à-dire, le nombre de fois où les articles i et j sont cités ensemble par les articles citants. $C(m, m)$ est une matrice carrée symétrique qui par construction contient sur la diagonale les fréquences de citation des articles cités dans le corpus.

Remarquons que la matrice de citation M n'est pas carrée ; elle ne représente pas un sous-graphe de citation, mais indique simplement quels sont les articles cités par l'ensemble citant. Ceci a une conséquence importante : la matrice de co-citation ainsi déduite ne contient pas toutes les relations de co-citation qui existent entre les articles cités mais seulement celles issues des articles citants (fig. 4.2). La figure 4.2 nous montre les relations de citation entre les sommets numérotés de 1 à 7. Si dans la cadre d'une étude bibliométrique, les éléments citants sélectionnés sont $\{1; 2; 4\}$, alors l'ensemble des éléments cités sera $\{4; 5; 6; 7\}$. Pour la construction de la matrice de co-citation entre les éléments cités, on ne tiendra compte que des citations émises par les éléments citants. Les citations issues des sommets 3 et 5 seront ignorées, et sur le graphe de co-citation, les relations entre les sommets 4 et 6 d'une part et 6 et 7 d'autre part, n'apparaîtront pas. Ceci est tout à fait volontaire : rappelons que l'objectif initial de la méthode des co-citations est la structuration du corpus citant. Elle utilise comme variable les citations (à la place des mots par exemple) et la structuration des articles cités n'est qu'une étape intermédiaire qui n'a de sens qu'au regard des articles citants.

Notre objectif est différent : nous voulons structurer un corpus issu du Web, en étudiant les *sitations* que les pages reçoivent, et en particulier les relations de *co-sitation* qu'elles entretiennent. Pour cela, nous devons obtenir pour chaque page de notre corpus ses « prédécesseurs », c'est-à-dire l'ensemble des pages pointant vers elle, les pages citantes. En 2001, le moteur de recherche

[1] `http://www.isinet.com/`

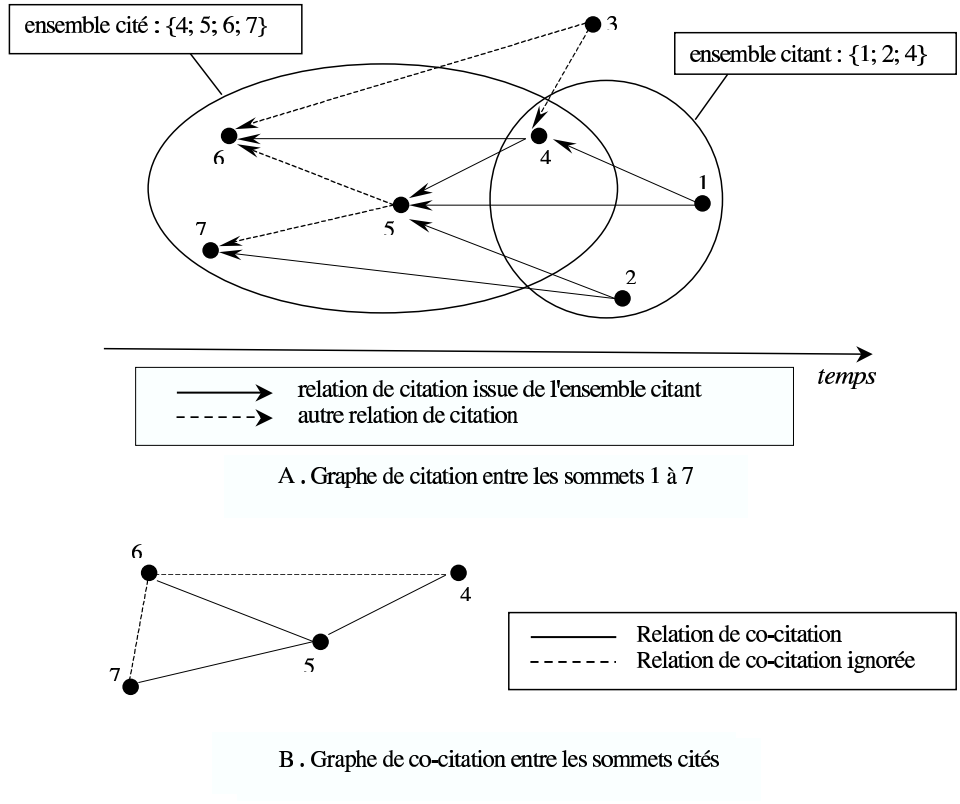

FIG. 4.2 – Construction du graphe de co-citation dans le cadre d'un étude bibliométrique traditionnelle

Google[2] offrait la possibilité de retrouver les prédécesseurs de chaque page, avec une fonction particulière nommée *link*.

Remarquons qu'une telle démarche n'est pas envisageable en bibliométrie classique, puisque le corpus citant à structurer est généralement récent et l'on ne peut pas prévoir les citations qu'il recevra. Le Web ne présentant pas de caractère diachronique, nous pouvons agir ainsi, avec une limite cependant : les URLs trop récentes ne sont pas ou faiblement citées, et ne pourront participer à la structuration.

La formation de notre corpus s'est déroulée en trois étapes que nous allons décrire (fig. 4.3). Précisons toutefois que les tâches d'interrogation et de rapatriement de données ont pu être automatisées par l'écriture de scripts en langage de programmation *Perl* et utilisant plus particulièrement la commande *Curl*[3] disponible sous différents *UNIX*.

[2] `http://www.google.com`
[3] `http://curl.haxx.se`

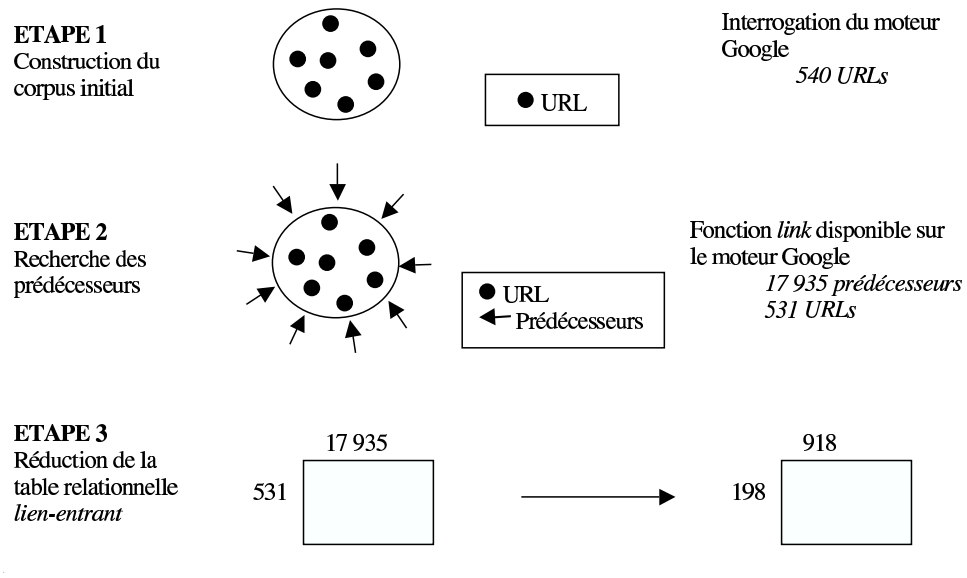

FIG. 4.3 – Etapes de la formation de notre corpus

4.3.2 Etape 1 : Construction du corpus initial

Au cours du mois d'août 2001 nous avons interrogé le moteur de recherche Google avec la requête « astronomie » réduite aux pages françaises. Ce moteur donnait comme estimation 54.000 réponses mais n'affichait qu'une liste de 540 URLs considérées comme les plus pertinentes. Notre robot nous a permis d'obtenir la liste de ces 540 URLs que nous avons sauvegardée dans un fichier nommé *corpus-initial*. Le faible taux de pages retrouvées par rapport à l'estimation donnée par Google ne constitue pas une limite pour notre travail. En effet notre ambition n'est pas de structurer l'ensemble des pages discutant d'astronomie, mais d'organiser un corpus en utilisant ses liens de *co-sitation*. Le tableau 4.1 rappelle les caractéristiques du fichier *corpus-initial*. Le tableau 4.2 présente la distribution du nombre de pages retrouvées pour chaque site de *corpus-initial*.

Etape 1 : Interrogation du moteur Google avec la requête « astronomie » en français				
Nom du fichier	Champs	Nombre de lignes	Nombre d'URLs distinctes	Nombre de sites distincts
corpus-initial	URL	540	540	424

TAB. 4.1 – Caractéristiques du fichier *corpus-initial*

Nombre de pages retrouvées par site	Nombre de sites
1	325
2	82
3	17
total	424

TAB. 4.2 – Distribution du nombre de pages retrouvées par site

4.3.3 Etape 2 : Recherche des prédécesseurs

La seconde étape (fig. 4.4)[4] est la recherche des prédécesseurs pour chacune des 540 URLs de notre corpus. Les prédécesseurs s'obtiennent en interrogeant successivement le moteur Google pour chaque URL avec la requête *link* appropriée. Les résultats sont sauvegardés dans un fichier nommé *lien-entrant*. Ce fichier est une table relationnelle composée de trois champs :

- URL : qui correspond à une adresse URL du fichier *corpus-initial*
- $Pred$: qui correspond à l'adresse d'un des prédécesseurs de URL
- $Type$: qui donne la nature du lien entre URL et $Pred$. Le lien est qualifié d'interne si URL et $Pred$ sont hébergés sur le même site et externe dans le cas contraire[5].

Notre fichier *lien-entrant* (Tab. 4.3) contient 21.739 relations de citations, 17.935 prédécesseurs distincts (pages citantes) et 531 URLs de notre corpus initial, ce qui signifie que pour 9 URLs aucun prédécesseur n'a été retrouvé. Cette perte d'URLs s'explique par le paramétrage de notre programme qui impose un temps maximum pour rapatrier les données (*timeout* fixé à deux minutes), au delà de ce temps les requêtes sont interrompues.

Etape 2 : Recherche des prédécesseurs					
Nom du fichier	Champs	Nombre de lignes	Nombre d'URLs distinctes	Nombre de sites distincts	Nombre de prédécesseurs distincts
lien-entrant	URL Pred Type	21.739	531	416	17.935

TAB. 4.3 – Caractéristiques du fichier *lien-entrant*

D'autre part, nous savons aussi que la recherche des prédécesseurs est incomplète. En effet, il existe une volonté de la part des concepteurs des moteurs

[4]La figure 4.4 illustre les relations de citation (internes et externes) des prédécesseurs vers les URLs. Il existe une intersection entre les prédécesseurs et les URLs du *corpus-initial* dont on ne s'occupe pas dans le cadre de cette expérience.

[5]Pour définir un site nous utilisons la seconde approximation présentée dans section 3.2.3, c'est-à-dire la concaténation d'un nom de machine et d'un nom de domaine.

de ne pas dévoiler la totalité de leurs informations. Nous l'avions déjà remarqué pour la requête « astronomie » où Google indique « environ 54.000 réponses » mais n'en affiche que 540. Dans une étude menée sur plusieurs moteurs, Bar-Ilan montre que cette rétention d'information est particulièrement vraie pour les requêtes utilisant la fonction *link* [Bar-Ilan, 2001]. Il est donc impossible d'obtenir avec cet outil un sous-graphe exhaustif du graphe web et ceci représente une limite pour notre expérimentation.

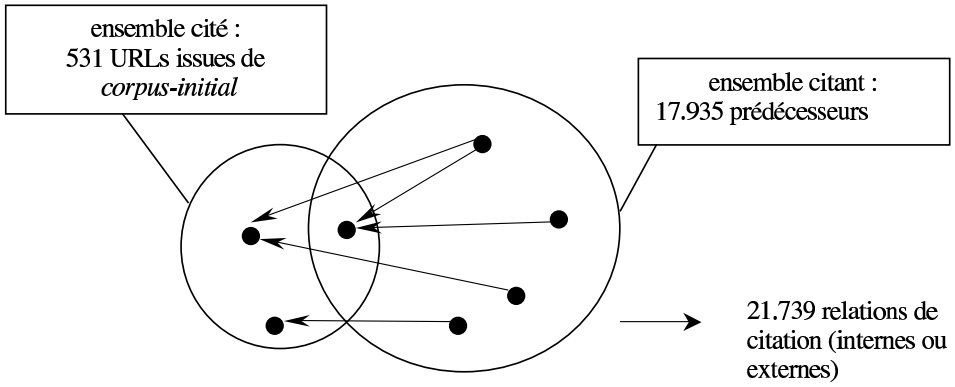

FIG. 4.4 – Représentation de la table lien-entrant

Notre table relationnelle *lien-entrant* peut s'écrire sous la forme d'une matrice de citation (17.935×531). Nous insistons sur le fait que cette matrice n'est pas un sous-graphe du Web. La distribution des citations émises n'a aucun sens ici, puisque que nous n'avons que les citations émises vers les pages de notre corpus initial. Par contre, la distribution des citations reçues devrait être significative et comparable aux résultats de la littérature malgré les limites évoquées ci-dessus. La figure 4.5 présente la distribution des citations reçues par les URLs de notre corpus (degrés entrants). Cette courbe en échelle log-log montre que la distribution est hyperbolique, moins dispersée que celle obtenue par Broder et al. [Broder et al., 2000] (figure 2.12, page 44). Ceci s'explique facilement au vu de la faible taille du corpus et de son mode de formation : nous ne devons pas ignorer que les pages rendues par Google ont en principe un bon PageRank et donc un degré entrant important.

4.3.4 Etape 3 : Réduction de la Table relationnelle *lien-entrant*

Comme nous l'avons évoqué au cours du chapitre précédent (cf. chapitre 3, section 3.2.1) les liens intra-sites (liens internes) sont analogues à l'auto-citation et servent surtout à la navigation dans les sites. Finalement, ce sont plutôt les liens inter-sites (liens externes ou *sitations*) qui matérialisent des relations remarquables entre pages citantes et citées. C'est pourquoi, notre méthode de structuration basée sur la *co-sitation* n'utilise que les relations externes entre

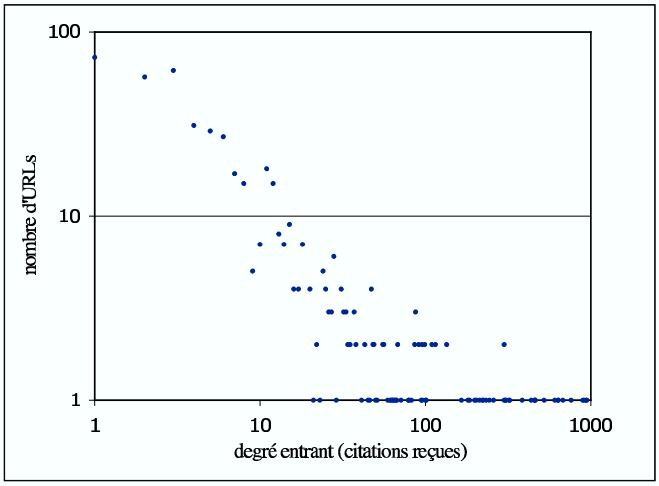

FIG. 4.5 – Distribution des citations reçues en échelle log/log

les éléments citants et cités. Nous supprimons toutes les relations internes présentes dans la table *lien-entrant*, c'est-à-dire 8.547 relations de citation. Nous obtenons une nouvelle table relationnelle nommée *lien-entrant-ext* composée de $21.739 - 8.547 = 13.192$ lignes. La suppression des 8.547 relations entraîne une perte importante d'URLs : 228 URLs qui ne reçoivent que des liens de citation interne. Pourtant, nous ne devons pas nous alarmer devant cette réduction de corpus : ces 228 URLs correspondent aux pages qui ne reçoivent aucune citation externe (*sitation*) et qui par conséquent ne sont pas des points d'entrée.

D'autre part, pour être *co-sitée*, une page doit être citée par des pages émettant plusieurs *sitations*. Cette condition que nous avons déjà évoquée constitue certainement la limite la plus importante de la méthode, et nous consacrerons toute une partie du chapitre 6 pour essayer de l'évaluer. 10.475 prédécesseurs ne citent qu'une seule page de notre corpus et seront supprimés. Finalement, la table relationnelle sur laquelle nous allons travailler, en particulier pour construire le graphe de *co-sitation*, ne contient plus que 2.717 relations de *sitation*, 198 URLs et 918 prédécesseurs (pages citantes). Nous nommons cette table *relasit*, pour « relations de *sitation* ». Cette table peut s'écrire sous la forme d'une matrice de citation $M(918, 198)$. Cette matrice est binaire et très creuse. Le tableau 4.4 résume les différentes étapes de réduction de la table relationnelle de citation. Les 198 URLs de notre table *relasit* constituent le corpus de test sur lequel nous allons mener notre expérimentation. La figure 4.6 nous montre la distribution des degrés entrants pour les pages de ce corpus. Sur cette figure, la droite d'équation y ($y = 39,877x^{-1,0126}$) est un ajustement assez significatif de cette distribution (le coefficient de détermination R est proche de $0,85$). Ceci montre une distribution moins dispersée que celle obtenue par Broder et al [Broder et al., 2000] et qui peut s'expliquer comme précédemment. D'autre part, cette distribution ne concerne que les *sitations* reçues (citations

Etape 3 : Réduction de la table relationnelle *lien-entrant*					
Nom du fichier	Champs	Nombre de lignes	Nombre d'URLs distinctes	Nombre de sites distincts	Nombre de prédécesseurs distincts
liens-entrant (rappel étape 2)	URL Pred Type	21.739	531	416	17.935
liens-entrant-ext	URL Pred Type = "Externe"	13.192	303	265	11.393
relasit	URL freq.prédécesseur > 1	2.717	198		918

TAB. 4.4 – Caractéristiques du fichier *lien-entrant-ext*

inter-sites) et par conséquent, elle n'est pas vraiment comparable aux résultats de la littérature.

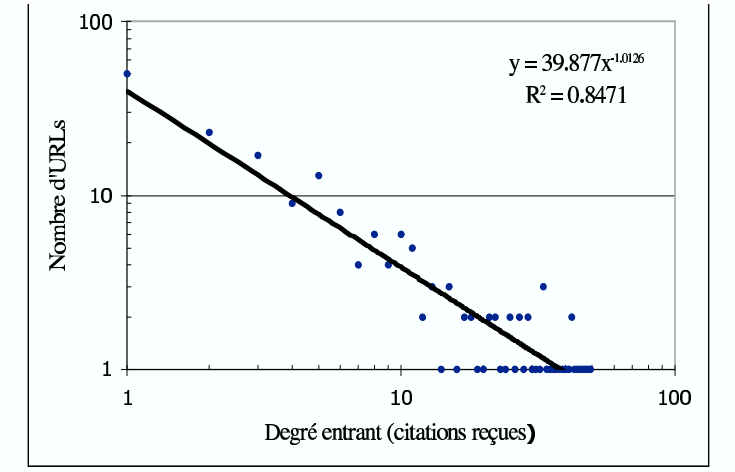

FIG. 4.6 – Distribution des degrés entrants dans le corpus final

4.4 Qualification manuelle du corpus de test

Dans un premier temps, nous avons procédé à la qualification manuelle des 198 pages de notre corpus pour les quatre métadonnées définies au chapitre précédent : *Type d'autorité*, *Type de site*, *Type de page*, *Type d'information*. Pour compléter cette qualification, nous avons défini des sous-classes pour les métadonnées *Type d'autorité* et *Type de site*. Ces sous-classes sont dépendantes du corpus de l'étude. Elles n'interviennent pas par la suite dans les calculs d'homogénéité, mais elles permettent de mieux comprendre l'apport informationnel de chaque page du corpus. Les sous-classes pour la métadonnée *Type d'autorité*

sont les suivantes :

- institution,
  - centre de recherche,
  - bibliothèque,
  - ministère,
  - enseignement (primaire, secondaire ou supérieur),
  - musée,
  - office du tourisme,
  - (structure d') animation scientifique,
- association,
  - club amateur,
  - (structure d') animation scientifique,
  - société (club professionel),
- personne individuelle,
- entreprise.

Les sous-classes pour la métadonnée *Type de sites* sont les suivantes :

- homeserveur,
- site de recherche,
  - annuaire,
    - (recherche de) pages,
    - (recherche de) logiciels,
    - (recherche de) services,
  - moteur,
    - (recherche de) pages,
    - (recherche de) logiciels,
    - (recherche de) services,
- site de ressources,
  - (contenant des) documents,
  - (contenant des) images,
  - (contenant du) son,
  - (contenant des) logiciels,
- service web.

Pour cette indexation, nous avons utilisé à la fois l'information contenue sur les 198 pages, ainsi que celle se trouvant sur le site dont elles font partie. Deux difficultés ont été rencontrées lors de l'indexation.

- Plusieurs fois, il s'est avéré impossible d'attribuer les valeurs de métadonnées. Non pas que la typologie soit imprécise ou incomplète, mais plutôt par manque d'information disponible sur les sites. Repérer qui est à l'origine du site par exemple, n'est pas toujours une tâche simple, en particulier pour les sites de ressources ou de recherche dont l'objectif premier n'est pas de présenter l'autorité. D'autre fois, c'est le rôle informationnel du site qui n'est pas clairement déterminé. Lorsque nous ne savions pas quelle valeur affecter, nous avons renseigné avec la mention « indéterminé ».

– La deuxième difficulté est apparue lorsqu'une page semblait appartenir à plusieurs catégories, en particulier pour la métadonnée *Type de site*. Certains sites proposent d'importantes bases de données concernant les produits ou les services de l'autorité qui les a créés (e.g. les bases de données des horaires de train disponibles sur le site de la SNCF). Ces sites sont considérés comme des homeserveurs et non comme des sites de ressources, car l'information proposée est relative à leur activité. Il existe toutefois une exception. Elle concerne les sites dont l'activité de l'initiateur est relative au traitement de l'information et à la communication. Ainsi, les sites d'éditeurs ou de journaux qui proposent en priorité et en grande majorité de l'information documentaire (accès aux articles, par exemple) sont considérés comme des sites de ressources. Voici quelques exemples de qualification de pages pour la métadonnée *Type de page* figurant dans notre corpus.
  * Notre corpus contient différents sites de librairies[6] proposant un accès aux catalogues. Ces sites sont qualifiés d'homeserveur, car ils ne proposent que des renseignements basiques sur les livres et aucune information de fonds.
  * Ce corpus comporte aussi des pages issues de sites de journaux[7]. Ces sites proposent de l'information élaborée (accès aux articles) et non auto-descriptive et sont qualifiés de sites de ressources.

Les résultats de l'indexation sont présentés en annexe (Annexe A). Le tableau 4.5 nous en donne le tri à plat. Ce corpus contient majoritairement des sites homeserveurs (63% des pages) et des pages d'accueil (67% des pages). Ceci confirme l'idée de H. Chu [Chu, 2004] selon laquelle se sont surtout les pages d'accueil qui sont citées.

4.5 Structuration du corpus par la méthode des *co-sitations*

Parallèlement à l'indexation, nous avons structuré notre corpus par la méthode des *co-sitations*. Seule la première étape de la méthode, celle consistant à regrouper les éléments les plus cités en agrégats (cf. 2, page 36) nous intéresse ici. Cette opération comporte trois phases. La première est le calcul de la matrice de *co-sitation* entre les URLs ; la seconde détermine la similarité entre les URLs basée sur la force de *co-sitation* ; enfin, la dernière est le regroupement de ces URLs en aggrégats utilisant des méthodes de classification automatique.

[6] URL 37 `http://www.blanchard75.fr/` ; URL 39 `http://www.burillier-uranie.com/` ; URL 88 `http://www.galaxidion.com/` (les résultats complets de la qualification sont proposés en annexe (Annexe A).

[7] URL 100 `http://www.humanite.presse.fr/journal/jour.html/` ; URL 116 `http://www.ladepeche.com/` ; URL 117 `http://www.lexpress.presse.fr/Express/Info/Sciences/` ; URL 118 `http://www.liberation.com/sciences/`

Type d'autorité	nombre de pages	Type de site	nombre de pages
association	58	homeserveur	125
entreprise	53	site de recherche	24
institution	37	site de ressources	41
personne	37	service web	3
indéterminé	13	indéterminé	5
Type de page	**nombre de pages**	**Type d'information**	**nombre de pages**
page d'accueil	134	autodescriptive	103
page de contenu	28	non autodescriptive	89
index	13	indéterminé	6
indéterminé	5		
portail	17		
formulaire	1		

TAB. 4.5 – Résultat de l'indexation

4.5.1 Calcul de la matrice de *co-sitation*

La matrice de *co-sitation* C s'obtient en multipliant la transposée de la matrice de *sitation* par la matrice de *sitation* elle-même comme précédemment (équation 4.1).

$$C(198, 198) = M^t(918, 198) \times M(918, 198) \tag{4.2}$$

Cette matrice C est carrée symétrique :
- C_{ij} est la fréquence (ou force) de *co-sitation* du couple d'URLs (i, j),
- $C_{ii} = C_i$ est l'occurence de *sitation* de l'URL i, c'est-à-dire le nombre de fois où l'URL i est citée par les 918 prédécesseurs (appartenant à un site distinct).

La matrice de *co-sitation* (sans la diagonale) est une représentation du graphe de *co-sitations*, graphe valué où les nœuds sont les URLs et les arcs les liens valués donnant la force de *co-sitation*[8] entre celles-ci.

Cette matrice est relativement creuse et son graphe associé a une faible *densité*. La densité mesure le degré de connectivité global d'un graphe. Elle est définie comme le rapport entre le nombre d'arcs du réseau et le nombre d'arcs possibles. Dans le cas d'un graphe non orienté, le nombre d'arcs possible est :

$$\frac{N(N-1)}{2}, \tag{4.3}$$

où N est l'ordre du graphe (198 pour cette étude).
Dans notre cas, le nombre d'arcs, c'est-à-dire le nombre de cases non nulles, sans

[8]Deux URLs i et j ont une force de *co-sitation* x, si elles sont citées x fois simultanément par les prédécesseurs.

	U18	U19	U20	U21	U22	U23	U24
U18	5	4	4	1	0	0	1
U19	4	6	4	0	0	0	0
U20	4	4	5	0	0	0	0
U21	1	0	0	14	0	0	1
U22	0	0	0	0	6	0	0
U23	0	0	0	0	0	2	0
U24	1	0	0	1	0	0	9

FIG. 4.7 – Extrait de la matrice de *co-sitation* entre les URLs 18 à 24

les cases de la diagonale, divisé par 2, est égal à 1701. La densité est donc

$$D = \frac{1701}{\frac{(198\times198)-198}{2}} = 0,087 \tag{4.4}$$

Ce graphe contient donc environ 9% des arcs possibles.

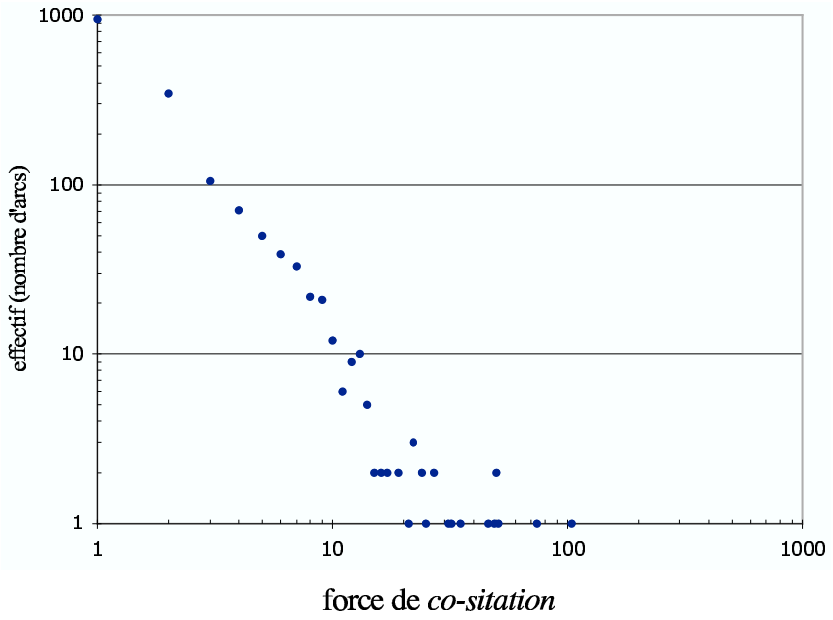

FIG. 4.8 – Distribution des forces de *co-sitation*

La figure 4.8 permet d'examiner la distribution des forces de *co-sitations*. Remarquons que celles-ci varient entre 1 et 104. Les six plus grandes valeurs étant : 104, 74, 51, 50, 49, 46. Nous nous intéressons aux couples d'URLs partageant les forces de *co-sitation* les plus élevées. Le tableau ci-dessous présente

ces couples par force de *co-sitation* décroissante. Le descriptif de l'information contenue sur les pages web correspondantes provient de l'indexation manuelle du corpus (cf. Annexe A).

Force de *co-sitations*	Numéro d'URL	URL	Descriptif
104	95 156	`http://www.globetrotter.net/astroccd` `http://www.quebectel.com/astroccd`	Groupe Astro et CDD Groupe Astro et CDD
74	102 103	`http://www.iap.fr/sf2a` `http://www.iap.fr/sfsa`	Société française d'astronomie et d'astrophysique Société française d'astronomie et d'astrophysique
51	143 164	`http://www.oceanet.fr/associations/san` `http://www.san-fr.com`	Société d'astronomie de Nantes Société d'astronomie de Nantes
50	164 142	`http://www.san-fr.com` `http://www.oceanet.fr/Associations/san`	Société d'astronomie de Nantes Société d'astronomie de Nantes
50	143 142	`http://www.oceanet.fr/associations/san` `http://www.oceanet.fr/Associations/san`	Société d'astronomie de Nantes Société d'astronomie de Nantes
49	172 189	`http://www.ecila.fr/french` `http://www.yahoo.fr`	Ecila France Yahoo France

Après consultation des fichiers sources des URLs du tableau, nous pouvons dire que les 5 premières valeurs sont partagées par des couples d'URLs « alias », c'est-à-dire des URLs pointant vers le même fichier de données, donc vers des pages web de même contenu. Il nous parait insensé qu'il existe de nombreuses pages citant simultanément des URLs « alias » créant ainsi des *co-sitations* d'URLs « alias ». C'est pourquoi, nous avons vérifié le code source de certains prédécesseurs de ces couples d'URLs : aucun de ceux-ci ne cite simultanément des URLs « alias ». Alors, comment expliquer la présence de ce phénomène

dans notre corpus ? Nous pensons plutôt que le moteur Google sait indentifier les URLs « alias », et que par conséquent, il leur retourne les mêmes réponses lors de la recherche des prédécesseurs. Notons qu'à partir de la sixième valeur, le problème des alias s'atténue et l'on voit apparaître des couples de pages réellement co-citées.

En co-citation classique, les graphes sont généralement plus denses et la structuration se fait en sélectionnant les paires d'articles ou d'auteurs les plus co-cités. Le seuil de co-citation se situe souvent autour de 5. Cette sélection permet d'une part, de réduire la taille de la matrice de co-citations en vue d'appliquer les algorithmes de structuration, et d'autre part, d'éliminer le bruit dû aux co-citations « anecdotiques »[9]. Le tableau ci-dessous montre le nombre d'URLs sélectionnées en fonction de la valeur du seuil.

seuil de *co-sitation*	1	2	3	4	5	6	7
Nombre d'URLs par seuil	198	132	102	87	77	60	56

Dans notre cas, la taille de la matrice est suffisamment petite pour appliquer sans difficulté les algorithmes de structuration sur la totalité des URLs. De plus, la présence de *co-sitations* « anecdotiques » est sans doute plus rare que dans le cas classique, puisque toutes les URLs à classer (éléments cités) partagent un point commun : le fichier vers lequel elles pointent contient le terme « astronomie ». Par conséquent, nous n'appliquons pas de seuil de *co-sitation* et gardons la totalité de nos URLs pour la structuration de notre corpus. Nous évitons ainsi une nouvelle réduction de corpus.

4.5.2 Calcul de la similarité entre les URLs

Pour déterminer la proximité entre les URLs, nous utilisons un *indice de similarité* qui traduit mathématiquement l'idée suivante : « deux URLs sont proches, si par rapport à leurs fréquences de *sitation* respectives, leur fréquence de *co-sitation* est importante ». Un indice de similarité est une application S de $E \times E$ vérifiant (E étant l'ensemble des éléments à classer, ici les URLs) :

$$\begin{cases} (a)\ S(i,j) = S(j,i) \\ (b)\ S(i,j) \geq 0 \\ (c)\ S(i,i) \geq S(i,j) \end{cases} \tag{4.5}$$

Il existe plusieurs indices possibles, qui comme la distance, sont symétriques (a) et positifs (b). Par convention, ils varient de 0 à 1 : ils sont égaux à 1 lorsque

[9] En bibliométrie classique, les articles scientifiques contiennent en moyenne entre 20 et 50 références ; tous les couples de références (éléments cités) ne traduisent pas forcément une association intéressante

les deux éléments apparaissent toujours ensemble, et à 0, lorsque ceux-ci n'apparaissent jamais ensemble. Contrairement à la distance, ces indices ne vérifient pas l'inégalité triangulaire et varient en sens inverse (c). Parmi les nombreux indices présentés et discutés dans la thèse de Bertrand Michelet [Michelet, 1988], nous avons choisi celui le plus fréquemment utilisé en scientométrie, notamment dans l'analyse des mots-associés, *l'indice d'équivalence* :

$$E_{ij} = \frac{C_{ij}^2}{C_i \times C_j} \in [0, 1]. \tag{4.6}$$

L'indice d'équivalence est bien un indice de similarité. En effet, il vérifie les axiomes :

- (a) : $C_{ij}^2 = C_{ji}^2$, donc $E_{ij} = E_{ji}$,
- (b) : C_{ij}^2, C_i, C_j sont positifs, donc $E_{ij} \geq 0$,
- (c) :
  (i) $E_{ii} = 1$
  (ii) $C_{ij} \leq C_i$ et $C_{ij} \leq C_j$, donc $\frac{C_{ij}}{C_i} \leq 1$ et $\frac{C_{ij}}{C_j} \leq 1$, donc $\frac{C_{ij}^2}{C_i \times C_j} \leq 1$
  d'après (i) et (ii) on peut déduire que $E_{ii} \geq E_{ij}$.

Cet indice est reconnu pour ses bonnes propriétés. Plus particulièrement, il s'agit d'un indice local. Un indice local ne fait pas intervenir le nombre N d'enregistrements. L'ajout d'enregistrements (dans notre cas, des prédécesseurs) ne contenant pas les éléments i et j ne modifie pas la similarité entre ces deux éléments.

Dans certains cas, lorsque le comportement citationiste des éléments citants est très disparate, il est préférable de pondérer l'indice choisi en fonction du nombre de liens émis (compte fractionnaire de citations). En effet, la *co-sitation* issue d'une page émettant 1.000 liens externes n'a sans doute pas la même valeur que celle issue d'une page émettant qu'un petit nombre de liens externes. De plus, les pages fortement citantes, comme les annuaires, multiplient les cooccurences non pertinentes de liens. Dans notre expérience, les prédécesseurs émettent entre 2 et 23 liens (externes) vers les URLs de notre corpus. Malgré la forte dispersion des *sitations* émises par les prédecesseurs (fig. 4.9[10]), nous choisissons de ne pas pondérer notre indice. En effet, nous ne connaissons pas le nombre de total de liens externes émis par les prédécesseurs mais seulement

[10]La droite y sur cette figure est un bon ajustement (le coefficient de détermination R est égal à 0,94). Bien que l'ensemble des liens sortants vers les 198 pages de notre corpus soit incomplet, le coefficient λ_s obtenu (2,67) est comparable aux résultats obtenus par Albert et al. [Albert et al., 1999] et Broder et al. [Broder et al., 2000].

la majorité des liens vers les URLs de notre corpus[11]. Dans ce cas, il apparaît difficile d'appliquer une méthode de pondération qui ne traduirait pas le comportement citationiste des prédécesseurs.

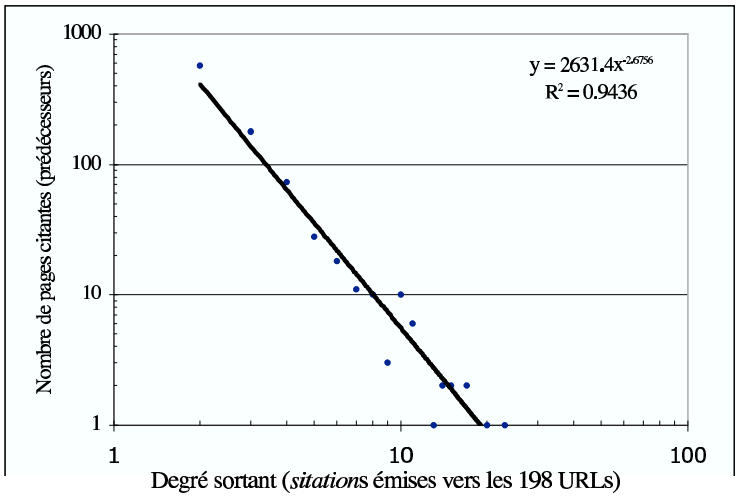

FIG. 4.9 – Distribution des degrés sortants pour les prédécesseurs

	U18	U19	U20	U21	U22	U23	U24
U18	-	0,53	0,64	0,01	0	0	0,02
U19	0,53	-	0,53	0	0	0	0
U20	0,64	0,53	-	0	0	0	0
U21	0,01	0	0	-	0	0	0,01
U22	0	0	0	0	-	0	0
U23	0	0	0	0	0	-	0
U24	0,02	0	0	0,01	0	0	-

FIG. 4.10 – Extrait de la matrice de similarité entre les URLs 18 à 24 avec l'indice d'équivalence

Les résultats de la proximité entre les URLs avec l'indice d'équivalence sont notés dans une matrice carré symétrique (fig. 4.10), qui est une représentation du graphe de *co-sitation*, à une différence près : les arcs valués entre les URLs ne correspondent pas à la force de *co-sitation* entre celles-ci, mais à leur similarité basée sur la *co-sitation*. Pour une meilleure lisibilité sur le graphe, il est possible de convertir l'indice de similarité en *indice de dissimilarité*. Un tel

[11] A titre d'exemple, le prédécesseur `http://fr.dir.yahoo.com/Sciences_et_technologies/Astronomie/Systeme_solaire` n'émet que 2 liens vers notre corpus, alors que nous savons que cette page contient de nombreux liens externes.

indice est une application vérifiant les propriétés suivantes :

$$\begin{cases} (a)\ D(i,j) = D(j,i) \\ (b)\ D(i,j) \geq 0 \\ (c)\ D(i,i) = 0 \end{cases} \tag{4.7}$$

Nous définissons alors $d_1(i;j)$ comme l'indice de dissimilarité associé à l'indice d'équivalence avec $d_1(i;j) = 1 - E(i;j)$. Cet indice varie comme une distance. Il prend la valeur nulle lorsque les éléments sont les plus proches possibles. La valeur 1 (valeur maximale) signifie qu'il n'existe pas d'arc (donc pas de *co-sitations* dans notre cas) entre les éléments.

La figure 4.11 présente l'extrait de la la matrice de dissimilarité valuée par l'indice de dissimilarité d_1 pour les URLs 18 à 24. La figure 4.12 représente le sous-graphe de *co-sitations* pour ces sept URLs.

	U18	U19	U20	U21	U22	U23	U24
U18	-	0,47	0,36	0,99	1	1	0,98
U19	0,47	-	0,47	1	1	1	1
U20	0,36	0,47	-	1	1	1	1
U21	0,99	1	1	-	1	1	0,99
U22	1	1	1	1	-	1	1
U23	1	1	1	1	1	-	1
U24	0,98	1	1	0,99	1	1	-

FIG. 4.11 – Extrait de la matrice de dissimilarité entre les URLs 18 à 24 avec l'indice d_1

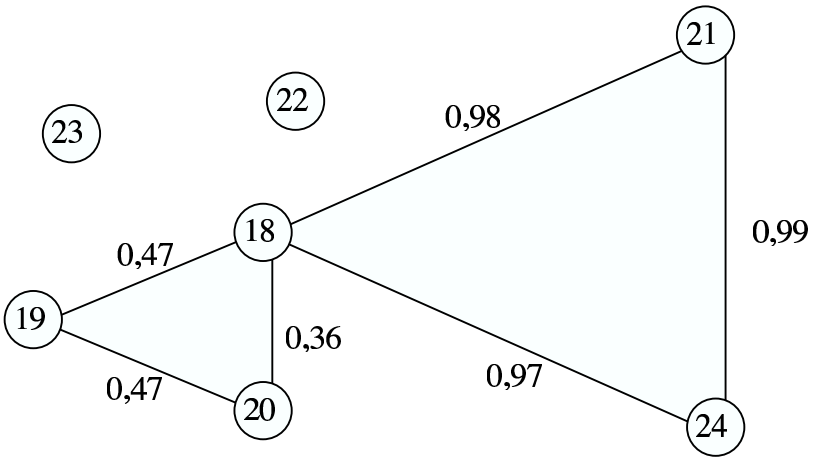

FIG. 4.12 – Extrait du graphe de *co-sitations* valué par l'indice de dissimilarité d_1

4.5.3 Regroupement des URLs en aggrégats

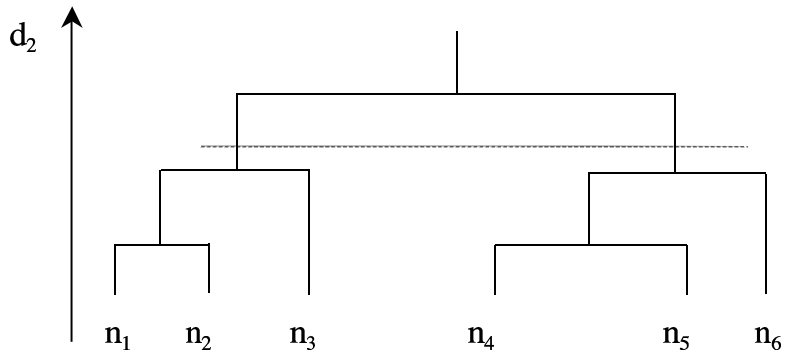

FIG. 4.13 – Exemple d'un dendrogramme

La troisième phase de la méthode est le découpage du graphe de *co-sitation* valué par l'indice de dissimilarité en vue d'obtenir des groupes homogènes (agrégats ou clusters). Ce découpage utilise une méthode de classification automatique provenant des statistiques multidimensionnelles [Benzecri, 1981], [Hartigan, 1975]. Plus particulièrement, nous avons utilisé une méthode de classification hiérarchique ascendante. Cette méthode consiste à fournir un ensemble de partitions du corpus. La méthode est dite *ascendante* car la première partition est constituée de chaque URL singleton ; les partitions intermédiaires sont constituées d'agrégats de plus en plus importants obtenus par regroupement successifs de parties ; la dernière partition rassemble toutes les URLs dans un seul et unique agrégat. On obtient alors une hiérarchie d'agrégats (disjoints) qui peut se représenter sous la forme d'un dendrogramme (ou arbre de classification) (fig. 4.13).

L'algorithme général peut être décrit de la manière suivante :

- étape initiale ($i = 0$) : associer un agrégat (cluster) singleton à chaque élément du corpus à classer ;
- soit G le nombre d'agrégats au niveau i.

Tant que G $>$ 1 étape i :
- balayer un tableau de $\frac{G(G-1)}{2}$ mesures de dissimilarité d_2 ;
- identifier les deux agrégats les plus proches et les regrouper en seul ;
- recalculer les mesures de dissimilarité d_2 entre les $G-1$ agrégats.

Fin Tant que

Au départ, les mesures de dissimilarité d_2 entre les agrégats singletons se lisent dans la matrice de dissimilarité calculée comme précédemment avec l'indice d_1 . Le problème ensuite est de définir la dissimilarité d_2 entre la réunion de deux agrégats et un troisième. Plusieurs stratégies sont possibles, les plus connues sont : le *simple lien* (ou saut minimum), le *lien moyen*, le *lien complet* [Turenne, 2000].

Dans la stratégie du simple lien, la dissimilarité entre la réunion de deux agrégats notés a et b avec un troisième noté c est :

$$d_2((a \cup b); c) = inf(d_2(a, c); d_2(b, c)). \tag{4.8}$$

La dissimilarité d_2 entre le nouvel agrégat $(a \cup b)$ et l'agrégat c est donc la plus petite dissimilarité d_2 entre a et c d'une part, et b et c d'autre part. Cette méthode découpe le graphe en agrégats peu denses et provoque des « effets de chaîne ». En effet, par cette méthode, les éléments de chaque agrégat sont proches d'au moins un autre élément de l'agrégat. Prenons comme exemple l'agrégat de la figure 4.14 comprenant les quatre éléments e_1, e_2, e_3, e_4. Dans cet exemple e_1 est proche de e_2, lui même proche e_3 qui lui-même est proche de e_4, sans que e_1 et e_3, e_1 et e_4, e_2 et e_4 ne soient directement reliés entre eux.

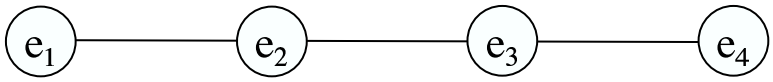

FIG. 4.14 – Exemple d'un effet de chaîne

A l'autre extrémité, dans la stratégie du lien complet, la dissimilarité d_2 entre le nouvel agrégat $(a \cup b)$ et l'agrégat c est la plus grande dissimilarité entre les parties :

$$d_2((a \cup b); c) = sup(d_2(a, c); d_2(b, c)). \tag{4.9}$$

Cette stratégie impose des relations entre tous les éléments de l'agrégat. En effet, pour qu'un nouvel élément en fasse partie, il faut que sa dissimilarité à tous les autres éléments soit inférieure au seuil. Les agrégats obtenus par cette méthode sont donc des cliques dans le graphe de *co-sitations*.

De nombreuses autres stratégies intermédiaires sont possibles. Elles définissent une ressemblance moyenne entre les agrégats. Dans la stratégie du lien moyen, la dissimilarité entre d_2 entre l'agrégat $(a \cup b)$ et l'agrégat c se calcule

comme suit :

$$d_2((a \cup b); c) = \frac{n_a}{n_a + n_b} d_2(a, c) + \frac{n_b}{n_a + n_b} d_2(b, c), \tag{4.10}$$

où n_a désigne le cardinal de l'agrégat a et n_b le cardinal de l'agrégat b.

La complexité de l'algorithme général décrit plus haut est en $O(n^3)$ en nombre d'opérations à effectuer, n étant le nombre d'objets à classer. Elle génère des arbres binaires composés de n partitions : à chaque fois que deux agrégats sont regroupés, la hiérarchie gagne un niveau auquel correspond une nouvelle partition. Chaque niveau est identifié par une valeur seuil qui est la valeur de dissimilarité à laquelle les deux derniers agrégats ont été regroupés. Un tel algorithme atteint rapidement les limites d'un ordinateur même puissant. C'est pourquoi, diverses implémentations ont été proposées pour accélérer le calcul des dendrogrammes. Nous avons utilisé une implémentation disponible au sein de notre laboratoire [Aguiar, 2002] qui requiert $O(n^2)$ en termes d'espace et de temps. La particularité de cette implémentation est de regrouper au même niveau les couples d'agrégats partageant les mêmes valeurs de dissimilarité. Les dendrogrammes ainsi obtenus ne sont pas binaires et comportent un nombre de niveaux nettement inférieur à n. Cette implémentation à été réalisée pour les trois stratégies mentionnées ci-dessus, ce qui a permis le découpage de notre graphe de trois manières différentes. Nous obtenons trois dendrogrammes composés de

- 115 seuils pour la stratégie du simple lien,
- 162 seuils pour la stratégie du lien moyen,
- 76 seuils pour la stratégie du lien complet.

Le point critique de toutes ces méthodes de classification est la détermination du niveau de coupure du dendrogramme qui donnera à la fois des clusters homogènes et de tailles importantes.

4.6 Analyse de l'homogénité des clusters

La phase finale de cette expérience consiste à croiser les résultats de l'indexation (section 4.4) avec ceux de la classification automatique (section 4.5). Notre objectif est l'analyse de l'homogénéité des agrégats pour un seuil de coupure et une stratégie donnés. Nous voulons examiner si les agrégats sont composés de pages partageant les mêmes valeurs de métadonnées. Cette section comporte trois points clés :

- la présentation des notions d'entropie et de redondance qui permettent de mesurer l'homogénéité des agrégats ;
- la détermination des métadonnées pour lesquelles la classification organise effectivement le corpus en agrégats homogènes ;
- l'étude globale de l'homogénéité des agrégats pour plusieurs métadonnées.

4.6.1 Notions d'entropie et de redondance

Lorsque l'on travaille avec des variables numériques, les notions de variance ou d'écart type permettent de mesurer la répartition des individus autour de la moyenne. Si l'on utilise des variables nominales, comme des métadonnées, la diversité des systèmes peut se mesurer en utilisant *l'entropie de l'information*. Cette notion introduite par Shannon en 1948 [Shannon, 1948] permet d'évaluer l'apport informationnel d'un système. « Plus un système est composé d'un grand nombre d'éléments différents, plus sa quantité d'information est grande, car plus grande est son improbabilité de le constituer tel qu'il est en assemblant au hasard ses constituants » [Atlan, 1979]. L'entropie d'un système se calcule par la formule suivante :

$$H = -\sum_{i=1}^{S} \frac{N_i}{N} \ln \frac{N_i}{N} \tag{4.11}$$

où
- N est le nombre d'éléments du système,
- S le nombre de modalités (valeurs différentes) que peuvent prendre les éléments,
- N_i l'effectif de chaque modalité.

Notons que l'entropie est nulle lorsque tous les éléments prennent la même valeur, c'est-à-dire lorsqu'une seule modalité est représentée dans le système. Dans ce cas là, il existe un $j \in [1, S]$ tel que $N_j = N$, et quelque soit $i \in [1, S]$ avec $i \neq j$, $N_i = 0$. L'entropie est maximale lorsque le système est le plus varié possible, c'est-à-dire lorsque toutes les modalités sont quasiment équitablement représentées, c'est-à-dire, quelque soit $i \in [1, S]$, N_i tend vers $\frac{N}{S}$.

La table 4.6 donne l'exemple d'un agrégat composé de 6 URLs et montre les valeurs de métadonnées obtenues par l'indexation manuelle. Appliqué à cet exemple :
- N est égal à 6 (nombre d'URLs),
- S est dépendant de la métadonnée étudiée et correspond au nombre de valeurs possibles pour la métadonnée choisie : par exemple pour la métadonnée *Type d'autorité*, S est égal à 4 (sans compter la valeur indéterminée),
- N_i est le nombre d'URLs décrites par chaque valeur de la métadonnée choisie : pour la métadonnée *Type d'autorité* les valeurs de N_i sont 0, 1, 1, 4 (cf. Tab. 4.6).

L'entropie de cet agrégat pour la métadonnée *Type d'autorité* est :

$$H = -\frac{2}{6} \ln \frac{1}{6} - \frac{4}{6} \ln \frac{4}{6} = 0,87$$

numéro d'URL	Type d'autorité	Type de site	Type de page	Type d'information	Descriptif
51	association	homeserveur	accueil	autodescriptive	Association utilisateurs de détecteurs électroniques.
194	institution	homeserveur	contenu	autodescriptive	Projet Céleste : contruction d'un détecteur au sol de rayon gamma.
121	entreprise	homeserveur	accueil	autodescriptive	Maison de l'astronomie : matériel.
18	entreprise	homeserveur	accueil	autodescriptive	L'Astronome : Société de matériel d'astronomie (grandes marques de matériel d'astronomie)
19	Entreprise	homeserveur	Accueil	autodescriptive	Paralux, société matériel astronomie. Spécialiste dans le domaine de l'optique instrumentale. Fabrication et commercialisation des jumelles, des instruments d'astronomie et des microscopes.
20	Entreprise	homeserveur	Accueil	autodescriptive	Société Astronomix : Société de matériel d'astronomie.

TAB. 4.6 – Exemple d'un agrégat

La *redondance* [Margalef, 1958] normalise cette fonction d'entropie. Elle varie entre 0 à 1 et mesure l'ordre d'un système plutôt que son désordre. Elle se calcule de la manière suivante :

$$R = \frac{H_{max} - H}{H_{max} - H_{min}}. \tag{4.12}$$

La redondance est égale à 1 lorsque l'entropie du système est minimum (le système est le plus ordonné possible) et nulle lorsque que H est maximum. Pour les distributions discrètes on a :

$$\lim_{N \to +\infty} H_{max} = \ln(S). \tag{4.13}$$

En effet, l'entropie d'un système est maximum lorsque chaque valeur est équitablement représentée, et lorsque N est grand, N_i tend vers $\frac{N}{S}$.

Pour N petit, il est préférable de ne pas utiliser cette approximation et de calculer les véritables valeurs de N_i. Soit q le quotient et r le reste de la division euclidienne de N par S. On a donc (fig. 4.15) r valeurs d'effectif $(q+1)$ et $(S-r)$ valeurs d'effectif q. H_{max} se calcule alors de la manière suivante :

$$H_{max} = -r\frac{q+1}{N}\ln(\frac{q+1}{N}) - (S-r)\frac{q}{N}\ln(\frac{q}{N}). \tag{4.14}$$

Dans notre exemple, $H_{min} = 0$ et H_{max} est obtenue lorsque les valeurs de métadonnées sont réparties de la manière suivante : 1, 1, 2, 2. H_{max} se calcule par l'équation 4.14 :

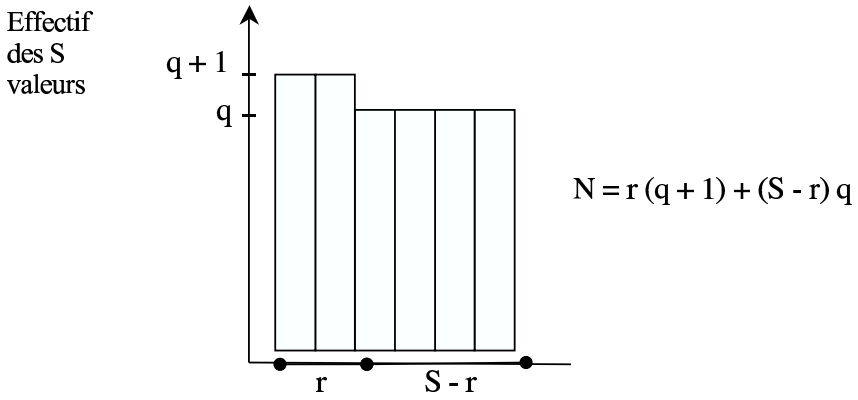

FIG. 4.15 – Histogramme des valeurs du système pour le cas H_{max}

$$H_{max} = -2 \times \frac{2}{6} \ln(\frac{2}{6}) - 2 \times \frac{1}{6} \ln(\frac{1}{6}) = 1,33.$$

La redondance de cet agrégat pour la métadonnée *Type d'autorité* se calcule conformément à l'équation 4.12 :

$$R = \frac{1,32 - 0,87}{1,32 - 0} = 0,35.$$

4.6.2 Etude du pouvoir organisateur de la classification métadonnée par métadonnée

Dans cette partie, nous allons étudier le pouvoir organisateur de la classification pour les quatre métadonnées *Type d'autorité*, *Type de site*, *Type de page* et *Type d'information*. Il s'agit de déterminer les métadonnées pour lesquelles la classification basée sur les liens découpe effectivement le corpus en agrégats homogènes. Cette étude se base sur un test, que nous avons mis au point et nommé *test d'homogénéité*, et sur une étude complémentaire tenant compte de la probabilité d'apparition de chaque agrégat.

4.6.2.1 Test d'homogénéité

Pour une métadonnée choisie, la redondance permet de mesurer l'ordre qui règne aussi bien :

- dans l'ensemble du corpus,
- que dans chaque agrégat.

Le test d'homogénéité que nous proposons permet de comparer pour une métadonnée les valeurs de redondance de chacun des agrégats avec celle du corpus total.

Les différentes métadonnées sont représentées par la variable λ qui prend les valeurs $\mathcal{A}$, $\mathcal{S}$, $\mathcal{P}$, $\mathcal{I}$ pour désigner respectivement les métadonnées *Type d'autorité*, *Type de site*, *Type de page* et *Type d'information*. Notons R^{λ}_{corpus} la valeur de redondance du corpus pour la métadonnée λ et R^{λ}_{a} est la valeur de redondance pour l'agrégat a.

Nous considérons qu'un agrégat est un regroupement ordonné de pages si sa valeur de redondance R^{λ}_{a} est significativement supérieure à celle du corpus (R^{λ}_{corpus}). La différence minimale que nous acceptons entre la valeur de redondance du corpus et celle d'un agrégat ordonné est fixée de manière empirique. Elle est notée ΔR^{λ}. Ce test est vrai si :

$$R^{\lambda}_{a} > R^{\lambda}_{corpus} + \Delta R^{\lambda}, \tag{4.15}$$

avec ΔR^{λ} compris entre $0 < \Delta R^{\lambda} < (1 - R^{\lambda}_{corpus})$.

La table 4.7 donne pour l'ensemble du corpus les valeurs d'entropie et de redondance pour nos quatre métadonnées. Pour ces calculs, les valeurs indéterminées ne sont pas prises en compte.

Métadonnée λ	nb valeurs	H	H_{max}	R^{λ}_{corpus}	$1 - R^{\lambda}_{corpus}$
Type d'autorité	4	1,365	1,386	0,015	0,985
Type de site	4	0,934	1,386	0,326	0,674
Type de page	5	0,956	1,609	0,406	0,594
Type d'information	2	0,690	0,693	0,004	0,996

TAB. 4.7 – Valeurs d'entropie et de redondance pour le corpus

Nous examinons pour ce test les résultats obtenus pour la méthode du lien complet au seuil de coupure le plus bas, c'est-à-dire avant que tous les agrégats ne soient regroupés ensemble. Ce dernier seuil est composé de 54 agrégats comportant 2 à 8 URLs et 38 singletons.

Nous avons calculé les valeurs de redondance pour l'ensemble des agrégats sans tenir compte des valeurs indéterminées. Nous procédons au test d'homogénéité en choisissant pour les quatre métadonnées $\Delta R^{\lambda} = 0,25$. La table 4.8 résume les résultats de ce test.

Les résultats sont particulièrement bons pour la métadonnée *Type d'information* avec 90 % des URLs appartiennant à des agrégats dont la redondance

	Type d'autorité		Type de site		Type d'information		Type de page	
	% d'agrégats	% d'URLs	% d'agrégats	% d'URLs	% d'agrégats	% d'URLs	% d'agrégats	% d'URLs
Test positif	77,08	84,83	70,59	77,78	86,27	90,85	58,82	65,36
$R_a = 1$	60,42	58,62	64,71	67,32	78,43	78,43	49,02	46,41
$R_a > 0,85$	62,50	64,14	66,67	72,55	80,39	83,66	49,02	46,41
Test négatif	22,92	15,17	29,41	22,22	13,73	9,15	41,18	34,64

TAB. 4.8 – Résultats du test d'homogénéité

est nettement supérieure à celle du corpus tout entier. De plus, 78% des pages appartiennent à des agrégats où l'homogénéité est maximale ($R_a = 1$).

Pour les métadonnées *Type d'autorité* et *Type de site*, les résultats sont bons, puisque que respectivement 84,83% et 77,78% des URLs sont issues d'agrégats respectant le test d'homogénéité.

Les résultats les plus décevants concernent la métadonnée *Type de page* où 40% des agrégats ne vérifient pas le test, ce qui correspond à 35% des URLs de notre corpus.

4.6.2.2 Etude complémentaire : probabilité de formation des agrégats

On considère que la classification découpe de manière ordonnée le graphe, si une majorité d'agrégats partagent des valeurs de redondance plus élevées que celle du corpus tout entier. Rappelons toutefois qu'au sein du corpus, les métadonnées ne sont pas distribuées de manière uniforme. Par exemple, la valeur *homeserveur* pour la métadonnée *Type de site* décrit 63% de pages de notre corpus. La probabilité de créer des agrégats dont toutes les pages partagent cette valeur est relativement forte, surtout si les agrégats sont petits. Des agrégats homogènes peuvent être formés de manière aléatoire, sans intervention particulière de la classification basée sur les liens. Pour s'assurer de la capacité réelle de notre méthode à ordonner les corpus, nous voulons étudier plus particulièrement les agrégats ayant une faible probabilité d'apparition. Ces agrégats sont-ils homogènes ?

Pour mener cette étude complémentaire, nous allons calculer pour chacun des agrégats sa probabilité de formation, et ce pour chacune des 4 métadonnées. Une telle probabilité se calcule par une loi hypergéométrique, loi basée sur la combinatoire.

Pour la métadonnée λ prenant S valeurs distinctes, N_i est le nombre d'URLs du **corpus** décrites par la i^{eme} valeur de cette métadonnée.
On a donc, $\sum_{i=1}^{S} N_i = N$, avec N le nombre d'URLs de notre corpus.

Soit un agrégat a composé de n URLs. Notons n_i le nombre d'URLs de l'**agrégat** décrites par la i^{eme} valeur de la métadonnée λ.
On a donc, $\sum_{i=1}^{S} n_i = n$.

La probabilité d'obtenir cet agrégat se calcule de la manière suivante :

$$P^{\lambda}(a) = \frac{C_{N_1}^{n_1} \times C_{N_2}^{n_2} \times ... \times C_{N_S}^{n_S}}{C_N^n} \tag{4.16}$$

La formule 4.16 nous permet le calcul de la probabilité d'apparition P de chaque agrégat. La figure 4.16 donne les quatre distributions (en rang décroissant) des probabilités d'apparition de chaque agrégat pour les métadonnées.

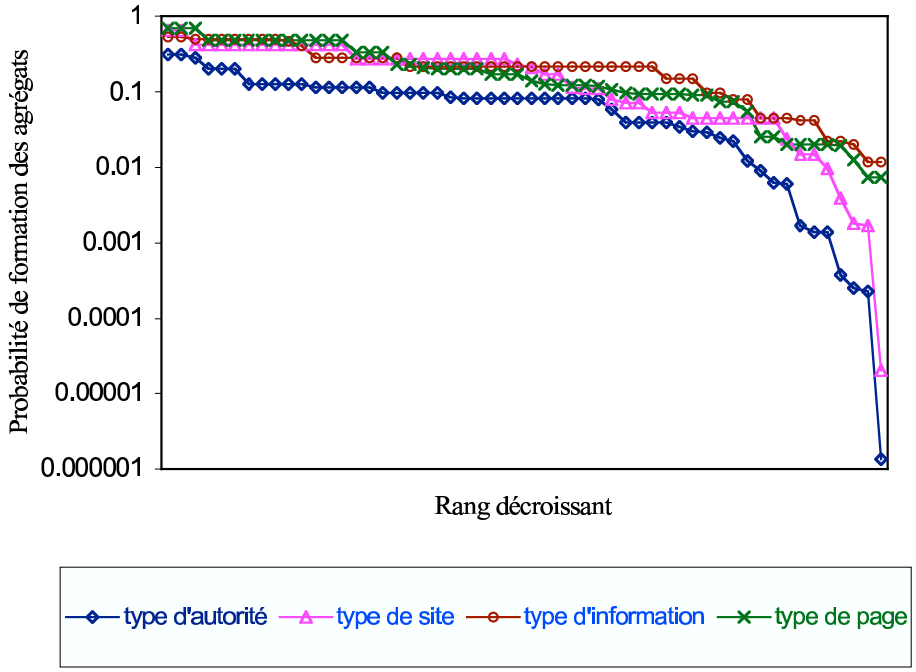

FIG. 4.16 – Distribution en rang décroissant des probabilités de formation des agrégats

Nous considérons de manière empirique qu'un agrégat a une faible probabilité de formation si celle-ci est inférieure à 5%. La table 4.9 donne pour les

	Type d'autorité		Type de site		Type d'information		Type de page	
	$P > 5\%$	$P \leq 5\%$	$P > 5\%$	$P \leq 5\%$	$P > 5\%$	$P \leq 5\%$	$P > 5\%$	$P \leq 5\%$
URLs (%)	40,00	60,00	68,59	31,41	63,46	36,54	85,62	14,38
agrégats (%)	58,33	41,67	76,47	29,41	81,48	18,52	80,39	18,87

TAB. 4.9 – Répartition des agrégats en fonction de leur probabilité d'apparition

Test homogénéité ($P \leq 5\%$)	Type d'autorité		Type de site		Type d'information		Type de page	
	négatif	positif	négatif	positif	négatif	positif	négatif	positif
URLs (%)	0	100	10,20	89,90	0	100	45,45	54,55
agrégats (%)	0	100	13,33	86,67	0	100	40	60

TAB. 4.10 – Résultats du test d'homogénéité pour les agrégats de faible probabilité d'apparition

différentes métadonnées la proportion d'agrégats et d'URLs en fonction de ce seuil. Dans cette étude complémentaire, ce qui nous intéresse tout particulièrement est l'étude de l'homogénéité des agrégats ayant une faible probabilité d'apparition ($P \leq 5\%$). La table 4.10 présente les résultats obtenus au test d'homogénéité (section 4.6.2.1) pour les agrégats ayant une faible probabilité d'apparition.

Ces résultats nous montrent que :
- pour les deux métadonnées *Type d'autorité* et *Type d'information*, tous les agrégats ayant une faible probabilité d'apparition sont nettement plus ordonnés que le corpus total,
- pour la métadonnée *Type de site*, les agrégats ayant une faible probabilité d'apparition respectent le test d'homogénéité à presque 90 %,
- pour la métadonnée *Type de page* les résultats ne sont pas convaincant puisque 45,5% des URLs appartiennent à des agrégats ne vérifiant pas le test.

4.6.2.3 Discussion

Les résultats du test d'homogénéité et de l'étude complémentaire montrent que la classification utilisant le principe de *co-sitation* permet de regrouper des URLs partageant une majorité de valeurs identiques pour les trois métadonnées *Type d'autorité*, *Type d'information* et *Type de site*. En ce qui concerne la quatrième métadonnée, *Type de page* les résultats ne sont pas concluants.

Le graphe web, et plus particulièrement le graphe de *co-sitation*, véhiculent bien de l'information liée à la typologie des sites et pages web. Les informations

véhiculées concernent plus le fond (l'autorité, le type de site, le type d'information) que la forme des documents. En effet, la forme des documents est représentée à travers la métadonnée *Type de page* qui n'a pas obtenu de bons résultats dans cette expérience. Ce résultat n'est finalement pas très étonnant. En effet, l'hypertextualisation produit des documents découpés en nœuds élémentaires hétérogènes. Les règles de découpage des documents en vue de produire des hypertextes ne sont pas normalisées. Ceci est particulièrement vrai dans l'univers du Web, si bien que la *co-sitation* de documents homogènes n'engendre pas forcément la *co-sitation* de nœuds (pages) se ressemblant physiquement.

4.6.3 Homogénéité globale des agrégats

Jusqu'à présent nous avons étudié les résultats obtenus indépendamment pour chaque métadonnée. Nous examinons à présent les résultats de chaque agrégat pour les trois métadonnées significatives dans cette expérience (*Type d'autorité*, *Type d'information* et *Type de site*).

Nous étudions les trois valeurs de redondance $R^{\mathcal{A}}$, $R^{\mathcal{S}}$, $R^{\mathcal{I}}$ obtenues par les 54 agrégats. Nous cherchons à savoir s'il existe une dépendance entre ces variables. Nous calculons les indices de corrélation pour les valeurs de redondance prises deux à deux. Voici les résultats obtenus :

- $Cor(R^{\mathcal{A}}; R^{\mathcal{S}}) = 0,101$,
- $Cor(R^{\mathcal{S}}; R^{\mathcal{I}}) = 0,303$,
- $Cor(R^{\mathcal{I}}; R^{\mathcal{A}}) = 0,183$.

Il n'existe aucune corrélation entre ces variables. L'ordre dans un agrégat pour une métadonnée, ne nous donne aucun renseignement sur l'ordre régnant dans cet agrégat pour les autres métadonnées.

Nous nous intéressons désormais à l'*ordre moyen* régnant dans chacun des agrégats pour les trois métadonnées significatives de l'expérience (*Type d'autorité*, *Type d'information* et *Type de site*). Pour un agrégat a, nous définissons l'ordre moyen $\overline{R_a}$ comme suit :

$$\overline{R_a} = moyenne(R_a^{\mathcal{A}}; R_a^{\mathcal{S}}; R_a^{\mathcal{I}}). \tag{4.17}$$

La table 4.11 résume ces résutlats. Cette table montre que 37% des URLs appartiennent à des agrégats où l'ordre est total pour les trois métadonnées, et que plus 60% des URLs appartiennent à des agrégats où la redondance moyenne est supérieure à 0,8. Les tables 4.12 et 4.13 présentant deux exemples d'agrégats où l'ordre est total.

De plus, l'examen manuel de notre corpus structuré (Annexe B) montre qu'il existe peu de clusters mélangeant à la fois de l'information autodescriptive et non-autodescriptive. En analysant précisement nos agrégats, nous constatons

Ordre moyen	nb d'agrégats	nb d'URLs	% cumulé
$\overline{R_a} = 1$	20	59	36,88%
$0,95 \leq \overline{R_a} < 1$	1	8	41,88%
$0,90 \leq \overline{R_a} < 0,95$	1	8	46,88%
$0,85 \leq \overline{R_a} < 0,90$	1	6	50,63%
$0,80 \leq \overline{R_a} < 0,85$	17	6	61,25%
$0,70 \leq \overline{R_a} < 0,80$	2	8	66,25%
$0,60 \leq \overline{R_a} < 0,70$	8	17	76,88%
$0,50 \leq \overline{R_a} < 0,60$	1	4	79,38%
$0,40 \leq \overline{R_a} < 0,50$	1	3	81,25%
$0,33 \leq \overline{R_a} < 0,40$	11	22	95%
$\overline{R_a} = 0$	4	8	100%
Total	54	160	

TAB. 4.11 – Distribution de l'ordre moyen régnant dans les agrégats

num URL	Type d'autorité	Type de site	Type d'information	Descriptif
27	association	homeserveur	autodescriptive	Association Astronomie Techniques et Communication
43	association	homeserveur	autodescriptive	Club d'astronomie Lyon ampère
181	association	homeserveur	autodescriptive	Club d'astronomie de l'université du Maine
30	association	homeserveur	autodescriptive	Club d'astronomiede l'école N7
124	association	homeserveur	autodescriptive	Association lunairienne d'astronomie
186	association	homeserveur	autodescriptive	Club d'astronomie
13	association	homeserveur	autodescriptive	Club d'astronomie

TAB. 4.12 – Exemple d'un agrégat où l'ordre est total

num URL	Type d'autorité	Type de site	Type d'information	Descriptif
67	entreprise	site de recherche	non autodescriptive	Annuaire de recherche généraliste
40	entreprise	site de recherche	non autodescriptive	Annuaire généraliste gratuit des sites francophones récents
53	entreprise	site de recherche	non autodescriptive	Minissimo, le guide web illusté des meilleurs sites internet
188	entreprise	site de recherche	non autodescriptive	Annuaire Whoyou
167	entreprise	site de recherche	non autodescriptive	Select link - annuaire francophone du Web gratuit

TAB. 4.13 – Exemple d'un agrégat où l'ordre est total

que lorsque les pages d'un agrégat contiennent de l'information autodescriptive et que celles-ci sont hébergées par des homeserveurs, l'autorité des pages est du même type (institutions, entreprises, etc.). Ainsi, nous avons plusieurs agrégats regroupant des centres de recherche, d'autres regroupant des club amateurs, etc. Par contre, lorsqu'un agrégat regroupe des pages pour lesquelles l'information est non-autodescriptive, le type d'autorité du site varie davantage. Ceci met en évidence que le Web est un lieu d'expression qui mêle à la fois des acteurs et des documents. En citant une page web décrivant l'autorité qui l'a créée, ce n'est pas l'apport documentaire de la page qui est pointé, mais plutôt son initiateur en temps qu'acteur, non pas du cybermonde, mais du monde réel [Rostaing et al., 1999]. L'étude du graphe par la méthode des *co-sitations* permet donc d'extraire des sous-corpus de documents relativement homogènes. Elle permet aussi d'identifier et de distinguer des réseaux d'acteurs et des réseaux de documents.

4.7 Discussion de l'expérience

Cette expérience visait l'extraction de sous-ensembles homogènes pour les métadonnées typologiques définies au chapitre 3. Cet objectif semble être atteint pour trois des quatre métadonnées étudiées grâce à notre structuration basée sur le principe de *co-sitation*. Cette structuration permet de mieux comprendre et appréhender l'univers du Web. En effet, elle organise les corpus et donne la possibilité de détecter à la fois des réseaux d'acteurs et des réseaux de documents. Elle ouvre des perspectives pour améliorer la recherche d'information, comme la qualification semi-automatique des pages par la propagation de métadonnées au sein des agrégats.
De plus, cette expérience montre aussi que la typologie proposée au chapitre 3 est adaptée aux ressources web, même si parfois nous ne disposons pas toujours de toute l'information nécessaire pour renseigner nos métadonnées.

Cependant, plusieurs limites ont été soulevées au cours de cette expérience. Les premières sont d'ordre technique et dépendent en partie des outils utilisés (manque d'information retournée par les moteurs de recherche). Nous sommes aussi confrontés à une limite théorique importante. Seule les URLs *co-sitées* peuvent être classées. Or nous savons qu'il existe de nombreuses pages sur le Web qui n'émettent qu'un seul lien. Toutes ces limites, plus la méthode de structuration utilisée qui génère des singletons au seuil de coupure, montrent qu'une grande partie des URLs ne peuvent pas faire partie de sous-corpus homogènes. Au cours du chapitre six, nous tenterons d'évaluer pour un corpus représentatif du Web français (5 millions de pages) la proportion de pages pouvant être classées par notre méthode.

Chapitre 5

Propagation de métadonnées

5.1 Introduction

L'affectation de métadonnées semble une bonne solution pour améliorer la recherche d'information sur la Toile. Comme nous l'avons mentionné au cours du premier chapitre, cette tâche est délicate et coûteuse en temps lorsqu'elle est effectuée manuellement. C'est pourquoi, les méthodes automatiques ou semi-automatiques apparaissent mieux adaptées pour affecter des métadonnées et ainsi décrire les ressources du Web. Les méthodes semi-automatiques offrent l'avantage de faire intervenir un jugement humain, ce qui permet d'enrichir le traitement effectué par les machines. Dans ce chapitre, nous présenterons deux méthodes semi-automatiques d'affectation de métadonnées basées sur la propagation : au départ, seule une partie des ressources sont qualifiées, leurs informations sont ensuite propagées aux autres ressources. Les méthodes de propagation proposées s'effectuent après la structuration des corpus par le principe de *co-sitation*.

L'idée de propager des métadonnées pour caractériser les pages web a été investie par Marchiori en 1998. Sa méthode [Marchiori, 1998] permet d'affecter des métadonnées de classification (thématique) et utilise la structure du graphe web (fig. 5.1). Dans sa méthode, les pages sont décrites par des métadonnées thématiques (mots-clés) pondérées par un coefficient variant entre 0 et 1 : la valeur 1 indique que la métadonnée décrit parfaitement la page ; la valeur 0 signifie qu'elle est inappropriée.

Son hypothèse est la suivante : si une page P (décrite par une métadonnée A pondérée par le coefficient ν) est citée par une page P', alors on peut supposer que P sert à expliciter (à appuyer) des idées évoquées dans la page P'. Les métadonnées de P peuvent donc être propagées « en arrière » à P' avec un facteur d'affaiblissement k ($0 < k < 1$). La métadonnée A décrit alors le

document P' avec le coefficient $\nu \times k$.

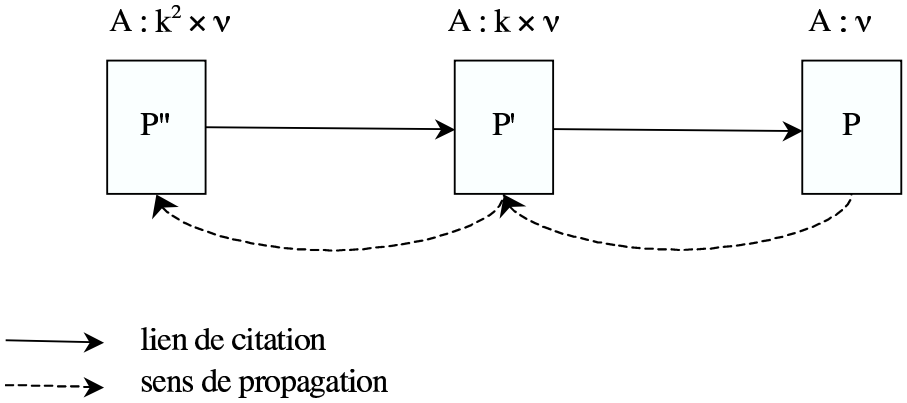

FIG. 5.1 – Propagation de métadonnées selon Marchiori

L'hypothèse de Marchiori stipule donc que deux pages reliées par un lien hypertexte partagent des métadonnées thématiques communes. Ceci est vrai pour les liens thématiques ou les liens cognitifs (cf. section 3.2.1, page 58), mais pas pour les autres types de liens (liens de navigation générale, liens gratuits, sociaux, etc.). De plus, cette hypothèse ne nous paraît pas valide pour d'autres métadonnées, telle que la métadonnée *Type de site*. Il n'y a a priori aucune raison pour que les sites citants et cités soient du même type. D'ailleurs très souvent, les pages hébergées sur des sites de recherche pointent vers des sites homeserveur ou des sites de ressources.

Comme Marchiori, nous pensons que le graphe du Web est porteur d'information et que celui-ci peut servir à l'affectation de métadonnées. Cependant, la relation de *co-sitation* nous paraît plus apte que la simple relation de *sitation* à rapprocher des pages partageant des propriétés communes (section 3, page 60). Les différents essais de transposition de la méthode de co-citation décrits dans la littérature, ainsi que les résultats du chapitre précédent, nous confortent dans cette idée. Ils montrent qu'une structuration basée sur le principe de *co-sitation* permet de regrouper des pages en agrégats homogènes, non seulement pour la dimension thématique des documents mais aussi pour leur dimension typologique.

5.2 Présentation des deux méthodes de propagation

Les deux méthodes développées dans ce chapitre reposent sur l'hypothèse H_2 évoquée au chapitre trois (section 3), que nous rapellons : Si une page P contient un lien hypertexte vers les pages P' et P'', il existe (au moins pour le

créateur de la page P) une raison pour citer ces deux pages ensemble. L'association existante entre les deux pages P' et P'' est d'autant plus forte, si elle est reprise par d'autres auteurs et si les pages P' et P'' sont toujours citées ensemble. Cette association se traduit par des valeurs identiques pour une ou plusieurs métadonnées.

Ces deux méthodes comportent trois étapes :

1. la structuration des corpus par le principe de *co-sitation* en vue d'obtenir une hiérarchie de sous-corpus que nous supposons homogènes,
2. l'affectation manuelle de métadonnées (indexation manuelle) pour un nombre limité de pages,
3. la propagation de métadonnées dans ces sous-corpus.

La première étape (la structuration des corpus par le principe de *co-sitation*) s'effectue de manière identique pour les deux méthodes et conformément à la méthode développée dans la section 4.5. Elle aboutit à la création d'un dendrogramme composé de $T+1$ niveaux. A chaque niveau t ($0 \leq t \leq T$ et $t \in \mathbb{N}$) correspond une valeur *seuil* (appartenant à $\mathbb{R}$) qui est la valeur de dissimilarité à laquelle les deux derniers agrégats ont été regroupés. Au niveau le plus bas ($t = 0$), la valeur seuil est nulle et chaque agrégat est un singleton, tandis qu'au niveau le plus haut ($t = T$), elle est égale à 1 et tous les éléments sont regroupés au sein d'un même agrégat.

Les deux méthodes différent ensuite par le choix des pages à indexer manuellement, et par la façon dont sont propagées les métadonnées. Toutes les deux utilisent la notion de distance dans les graphes. En effet, chaque agrégat induit un sous-graphe du graphe de *co-sitation*, dans lequel nous pensons que plus les éléments sont proches, plus grande est la probabilité qu'ils partagent des valeurs de métadonnée communes. Nous rappelons que dans un graphe valué, la distance géodésique d entre deux éléments est la somme des valuations du plus court chemin entre ceux-ci. Les chemins minimaux dans un graphe valué peuvent être identifiés par l'algorithme de Moore-Dijkstra que nous avons implémenté.

Algorithme de Moore-Dijkstra

L'algorithme de Moore-Dijkstra (1959) calcule les plus courts chemins d'un sommet à tous les autres sommets du graphe. Soit $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ un graphe valué composé de N sommets, muni d'une fonction $\gamma : \mathcal{A} \rightarrow \mathbb{R}_+^*$. Soit $\mathcal{N}(\mathcal{G}) = \{n_1; n_2; ...; n_N\}$ l'ensemble des N sommets de ce graphe, et $M(\mathcal{G})$ sa matrice d'adjacence associée. Rappelons que

$$M(i;j) = \begin{cases} 0 \; si \; i = j \\ \infty \; (ou \; l'ordre \; du \; graphe) \; si \; i \neq j \; et \; \{n_i; n_j\} \notin \mathcal{A} \\ \gamma(\{n_i; n_j\}) si \; i \neq j \; et \; \{n_i; n_j\} \in \mathcal{A} \end{cases} \tag{5.1}$$

Cet algorithme vise pour un sommet n_i $(1 < i < N)$, la construction par un procédé itératif d'un vecteur $\pi = (\pi(n_1), ..., \pi(n_j), ..., \pi(n_N))$ ayant N composantes, tel que $\pi(n_j)$ soit la longueur du plus court chemin (ou de la plus courte chaîne) allant du sommet n_i au sommet n_j. Le vecteur π est initialisé à $M(i, j)$, c'est à dire à la i^{eme} ligne de la matrice d'adjacence.
Cet algorithme considère deux ensembles de sommets :

1. S initialisé à $\{n_i\}$,
2. $\overline{S}$ son complémentaire.

A chaque pas de l'algorithme on ajoute un sommet à S jusqu'à ce que $S = \mathcal{N}(\mathcal{G})$. Voici la description de cet algorithme.

∗ **initialisation**
  · $\pi = M(i, j)$
  · $S = \{n_i\}$ et $\overline{S} = \mathcal{N}(\mathcal{G}) - \{n_i\}$
∗ **itérations**

Tant que $(\overline{S} \neq \emptyset)$
identifier un n_j dans $\overline{S}$ tel que $\pi(n_j)$ soit minimum,
retirer n_j de $\overline{S}$ et l'ajouter à S,

Pour (tout successeur n_k de n_j dans $\overline{S}$)
$\pi(n_k) \leftarrow \min\{\pi(n_k); \pi(n_k + \gamma(\{n_j; n_k\})\}$

Fin Pour

Fin Tant que

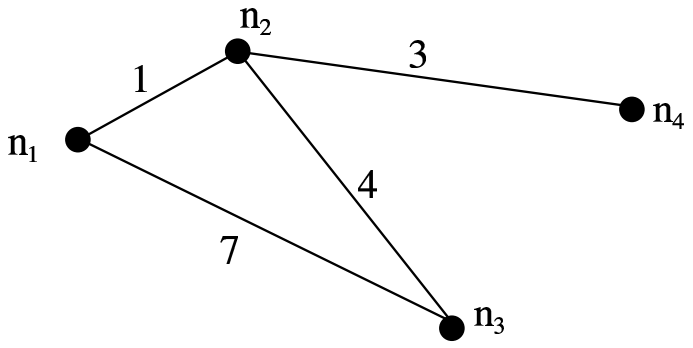

FIG. 5.2 – Exemple d'un graphe valué

Nous donnons comme exemple l'application de l'algorithme de Moore-Dijkstra au graphe valué représenté sur la figure 5.2. Nous cherchons à obtenir la longueur des plus courtes chaînes issues du sommet n_1.

– initialisation $S = \{n_1\}$; $\overline{S} = \{n_2; n_3; n_4\}$; $\pi = (1; 7; \infty)$
– première itération
  $n_j = n_2$ car $\pi(n_2) = 1 = \min\{1; 7; \infty\}$
  $S = \{n_1; n_2\}$; $\overline{S} = \{n_3; n_4\}$.

Les successeurs de n_2 dans $\overline{S}$ sont n_3 et n_4.
$\pi(n_3) \leftarrow \min\{7; 1+3\} = 4$
$\pi(n_4) \leftarrow \min\{\infty; 1+4\} = 5$
D'où le nouveau vecteur $\pi = \{1; 4; 5\}$.

– seconde itération $n_j = n_3$ car $\pi(n_3) = 3 = \min\{4; 5\}$
$S = \{n_1; n_2; n_3\}$; $\overline{S} = \{n_4\}$
n_3 n'a pas de successeur dans $\overline{S}$.
$\pi = \{1; 4; 5\}$

– troisième itération $n_j = n_4$
$S = \{n_1; n_2; n_3; n_4\}$; $\overline{S} = \{\emptyset\}$
$\pi = \{1; 4; 5\}$

Les longueurs des plus courtes chaînes entre le sommet n_1 et les sommets n_2, n_3, n_4 sont respectivement 1, 4 et 5.

5.3 Méthode 1

5.3.1 Présentation de la méthode

Dans les méthodes de classification, la détermination d'un niveau de coupure qui donnera à la fois des agrégats homogènes et de taille importante est un point critique. En effet, les études empiriques montrent que les agrégats ne se forment pas tous « à la même vitesse ». Certains sont déjà de taille importante et bien homogènes à un seuil relativement bas dans le dendogramme, alors que pour ce même seuil persistent encore beaucoup de singletons. A un seuil plus élevé, certains singletons ont pu se regrouper ou rejoindre d'autres agrégats pour former des ensembles homogènes, tandis que d'autres agrégats qui étaient homogènes se sont « bruités ». L'originalité de cette première méthode est d'utiliser la richesse de la hiérarchie : elle ne se limite pas à l'étude d'un seul niveau de coupure.

La méthode de propagation que nous proposons opère à partir d'un niveau K ($1 \leq K \leq T-1$), choisi de manière empirique. Il correspond intuitivement au niveau à partir duquel nous supposons que les agrégats sont déjà de taille importante et que ceux-ci ne sont pas encore trop bruités. Ce seuil dépend de la méthode d'agrégation choisie, plus la distance inter-agrégat d_2 est exigeante[1], plus ce seuil pourra être élevé. Cette méthode repose sur l'idée suivante : au sein d'un agrégat, si les pages les plus éloignées partagent les mêmes valeurs de métadonnées, alors les autres pages ont une forte probabilité de partager ces mêmes valeurs, puisqu'elles sont plus proches les unes des autres.

[1]La distance d_2 est définie section 4.5.3, page 89.

5.3.2 Algorithme de la méthode 1

L'agorithme que nous proposons pour cette première méthode est le suivant :

- étape initiale : $t = K$;
- soit π_t la partition d'agrégats obtenue au niveau t.

Tant que (il existe des pages non qualifiées dans le corpus) ET $(t > 0)$
étape t

Pour chaque agrégat de π_t composé d'au moins α pages, noté $Ag = \{P_i / 1 \leq i \leq n\}$ avec $n \geq \alpha$
Si $\forall i$, P_i est non qualifiée
- identifier les deux éléments les plus éloignés de l'agrégat[2] ;
- indexation manuelle des deux éléments les plus éloignés pour les métadonnées choisies ;

Si les valeurs de métadonnées sont identiques
propagation des valeurs aux autres éléments de l'agrégat
Fin Si
Fin Si
Fin Pour
$t \leftarrow t - 1$
Fin Tant que

5.3.3 Evaluation

L'objectif de cette section est d'évaluer notre méthode sur un corpus hypertexte provenant du Web. Il s'agit de comparer les résultats obtenus par notre méthode avec ceux d'une indexation manuelle. Malheureusement, nous ne connaissons pas de corpus de test comportant un sous-graphe du Web dont les pages sont qualifiées par des métadonnées. C'est pourquoi, nous choisissons de tester cette méthode sur le corpus formé dans le cadre de l'étude décrite au chapitre précédent. Rappelons que ce corpus contient 198 URLs, correspondant à des pages que nous avons indexées manuellement pour des métadonnées liées à la dimension typologique des documents. La structuration de ce corpus par les trois stratégies de classification, le simple lien, le lien moyen et le lien complet donne trois dendrogrammes comprenant respectivement 115, 162 et 76 niveaux (cf. section 4.5.3, page 89).

[2]Ce couple n'est pas forcément unique. Si plusieurs couples coexistent, un seul est choisi de manière aléatoire.

5.3.3.1 Propagation

Nous procédons à la propagation des valeurs métadonnées conformément l'algorithme décrit dans la section 5.3.2. La propagation des métadonnées a été réalisée pour les trois dendrogrammes, ainsi que pour tous les niveaux de départ K possibles. Les métadonnées choisies sont les métadonnées significatives de l'expérience précédente : *Type d'autorité*, *Type d'information* et *Type de site*. La taille minimum α des agrégats pour lesquels nous exécutons la méthode est fixée à 3. A chaque fois que la méthode doit faire intervenir un jugement humain, nous utilisons les résultats de l'indexation manuelle réalisée dans la section 4.4 (page 78). Au sein d'un agrégat, si les pages les plus éloignées sont indexées par des valeurs indéterminées, ces valeurs sont considérées comme différentes.

Notre implémentation contient deux modules :
- un module principal qui exécute l'algorithme de propagation (section 5.3.2) : il donne le nombre de valeurs de métadonnée propagées et le nombre de valeurs indexées manuellement ;
- un module comparant les résultats de la propagation avec ceux de la qualification manuelle : il donne le nombre de valeurs de métadonnées propagées identiques à celles obtenues manuellement (valeurs justes), le nombre de valeurs différentes (valeurs fausses) et le nombre de valeurs pour lesquelles nous ne pouvons conclure (cas où la valeur indexée manuellement est indéterminée).

Ainsi la méthode de propagation répartit les $3N$ valeurs de métadonnées en cinq cases (Tab. 5.1) :

Valeurs de métadonnée	Propagées	Non propagées
Justes	v_j^p	v_j^{non-p}
Fausses	v_f^p	v_f^{non-p}
sans conclusion	$v_?^p$	-

Tab. 5.1 – Répartition des valeurs de métadonnées

- $v_j^p + v_f^p + v_?^p$ est le nombre de valeurs de métadonnées propagées,
  - v_j^p est le nombre de valeurs de métadonnées propagées justes, c'est-à-dire identiques à celles obtenues par la qualification manuelle,
  - v_f^p est le nombre de valeurs fausses,
  - $v_?^p$ est le nombre de valeurs propagées pour lesquelles nous ne pouvons conclure.
- v_j^{non-p} est le nombre de valeurs de métadonnées qualifiées manuellement (valeurs des éléments les plus distants dans un agrégat).
- v_f^{non-p} est le nombre de valeurs de métadonnées non qualifiées (ni par propagation, ni manuellement). Il s'agit des valeurs décrivant les pages appartenant à des agrégats trop petits (nombre de pages inférieur à

α) pour lesquels la méthode n'est pas exécutée. Remarquons que $(v_j^p + v_f^p + v_?^p) + v_j^{non-p}$ correspond au nombre de valeurs de métadonnées qualifiées.

5.3.3.2 Méthode d'évaluation

Pour mesurer l'intérêt de la méthode de propagation, nous avons défini trois indices variant entre 0 et 1.

– La qualité de la propagation

$$Qual = \frac{v_j^p}{v_j^p + v_f^p} \in [0,1]. \tag{5.2}$$

C'est le rapport entre le nombre de valeurs propagées justes et le nombre de valeurs propagées. Cet indicateur mesure la précision de la propagation. Il reflète par ailleurs la cohésion au sein des clusters.

– La performance

$$Perf = \frac{v_j^p + v_f^p + v_?^p}{v_j^p + v_f^p + v_?^p + v_j^{non-p}} = \frac{v_j^p + v_f^p + v_?^p}{3N - v_f^{non-p}} \in [0,1]. \tag{5.3}$$

Elle donne une indication sur le nombre de valeurs de métadonnées propagées par rapport au nombre total de valeurs qualifiées manuellement et par propagation.

– Le taux de pages qualifiées

$$Taux = \frac{v_j^p + v_f^p + v_?^p + v_j^{non-p}}{3N} = \frac{3N - v_f^{non-p}}{3N} \in [0,1]. \tag{5.4}$$

Aux niveaux inférieurs de la classification, il existe beaucoup d'agrégats singleton ou d'agrégats de petite taille pour lesquels la méthode n'est pas appliquée ; par conséquent, de nombreuses pages ne sont pas qualifiées. Cet indice permet de savoir pour un niveau de départ donné (K), combien de pages sont qualifiées (de façon manuelle et par propagation) par rapport au nombre total de pages du corpus.

5.3.3.3 Présentation des résultats

Les résultats de cette expérimentation sont présentés sur les figures 5.3, 5.4 et 5.5. Plus particulièrement, la figure 5.3 s'intéresse aux valeurs de qualité obtenues par chaque niveau de départ K ($1 \leq K \leq T-1$) des trois stratégies de classification. Les trois graphiques de cette figure montrent que l'exigence de la méthode d'agrégation conditionne les résultats obtenus pour la qualité :

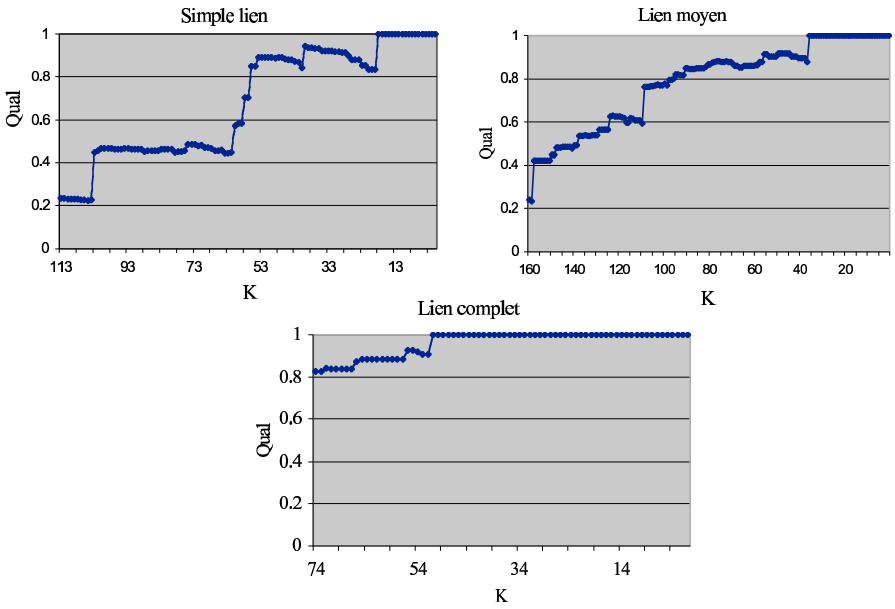

FIG. 5.3 – Représentation de la qualité pour les trois statégies (méthode 1)

- pour la stratégie du lien complet, les résultats sont les meilleurs. Ils varient entre 0,82 et 1. Globalement, la qualité croît régulièrement au fur et à mesure que le niveau de départ K décroît. A partir du niveau 51, c'est-à-dire pour 68% des niveaux de départ K, la méthode propage les informations sans erreur.
- pour la stratégie du lien moyen, les résultats varient entre 0,24 et 1. L'évolution de la qualité alterne des périodes à tendance d'accroissement avec des sauts positifs. Ces sauts correspondent aux niveaux pour lesquels la classification organise nettement le corpus. A partir du niveau 96, c'est-à-dire pour 60% des niveaux de départ K, la méthode propage les informations avec moins de 20% d'erreurs. A partir du niveau 36 (soit pour 22% des niveaux) la propagation s'effectue sans erreur.
- pour la stratégie du simple lien, les résultats varient eux aussi entre 0,24 et 1. L'évolution de la qualité alterne des phases plus ou moins stables (décroissant légèrement) avec des sauts. Les deux premières phases se situent autours des valeurs de qualité 0,22 et 0,45 ; elles concernent 45% des niveaux. A partir du niveau 56, c'est-à-dire pour 50% des niveaux de départ K, la méthode propage les informations avec moins de 20% d'erreurs, puis à partir du niveau 18 (soit pour 16% des niveaux) la propagation s'effectue sans erreur. Nous remarquons que pour les deux dernières phases, la qualité s'atténue niveau après niveau. En effet, la propagation se fait dans des agrégats relativement homogènes de plus en plus petits, mais où persistent toujours les mêmes erreurs. Ainsi pour le niveau 40, sur les 69 valeurs propagées, 65 sont justes et 4 sont fausses. Au niveau 19, seulement 24 valeurs sont propagées : 20 de ma-

nière correcte et toujours 4 avec erreur. C'est pourquoi nous observons une diminution de la qualité entre ces deux niveaux.

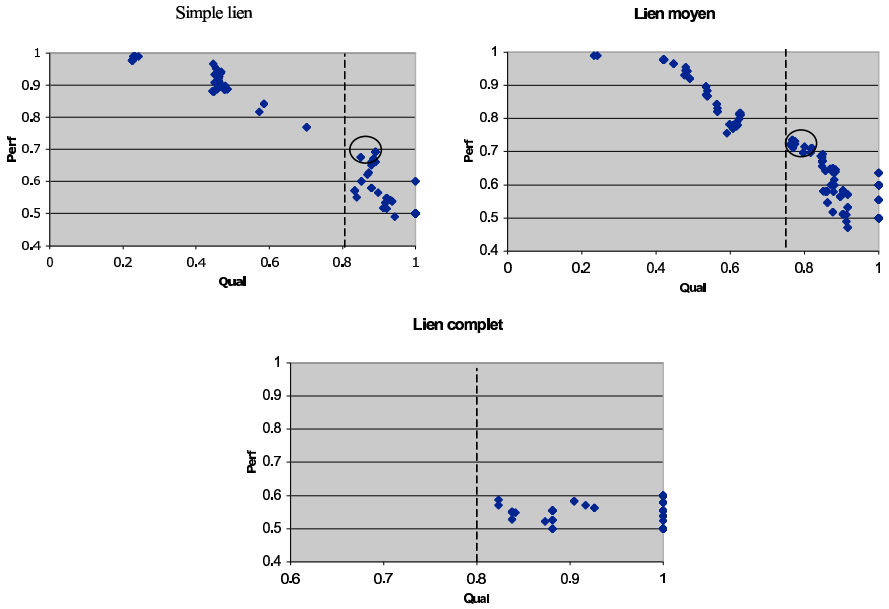

FIG. 5.4 – Représentation de la performance en fonction de la qualité pour les 3 stratégies (méthode 1)

Les valeurs de qualité obtenues pour cette méthode sont encourageantes. Quelle que soit la stratégie employée, la majorité des niveaux de départ obtient une qualité supérieure à 80% et de plus, il existe un niveau de départ K à partir duquel la propagation se fait sans erreur.

Nous étudions maintenant la relation entre la qualité et la performance. La figure 5.4 donne l'évolution de la performance $Perf$ en fonction de la qualité $Qual$ pour les trois stratégies.

- Pour les stratégies du simple lien et du lien moyen, les courbes présentent la même forme. Elles sont décroissantes et la performance varie entre 0,4 et 1. Ces courbes montrent une corrélation négative entre la qualité et la performance, surtout pour les valeurs de qualité faibles (inférieures à 0,8).
- pour la stratégie du lien complet, les valeurs de performance varient entre 0,5 et 0,6 ce qui reste faible. De plus, aucune corrélation n'apparaît.

Ces courbes montrent que pour une qualité convenable (supérieure à 0,80), les meilleurs résultats de la performance[3] atteignent au mieux 0,71 pour le lien

[3]Les meilleurs résultats de performance pour une qualité supérieure à 0,8 apparaissent entourés sur les graphiques du simple lien et du lien moyen (figure 5.4).

moyen (0,82 ; 0,71) et 0,69 pour le simple lien (0,89 ; 0,69). Appliquée à ces dendrogrammes, notre méthode ne paraît pas très rentable.

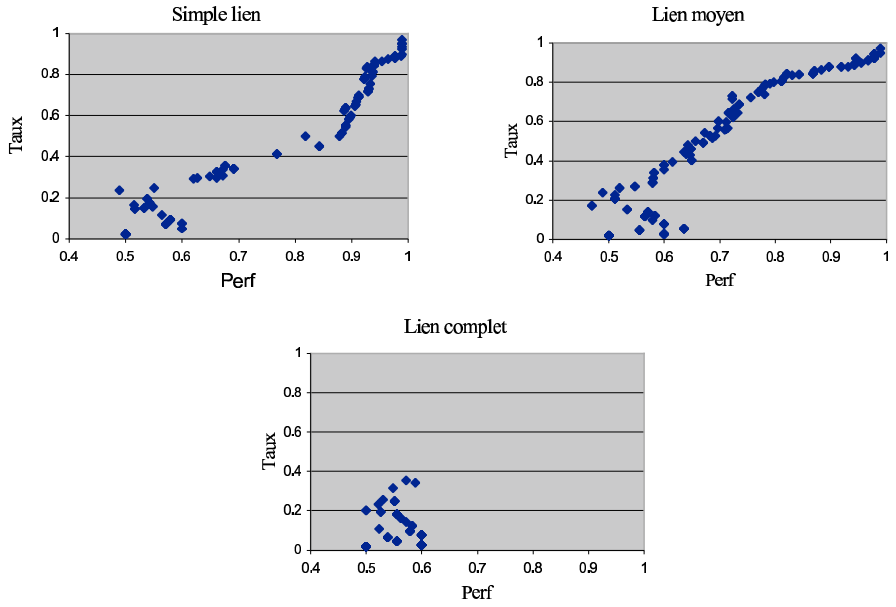

FIG. 5.5 – Représentation du taux de qualification en fonction de la performance pour les 3 stratégies (méthode 1)

La figure 5.5 montre l'évolution du taux de pages qualifiées en fonction de la performance.

- Pour les stratégies du simple lien et du lien moyen, les courbes présentent sensiblement la même forme. Elles sont croissantes et les taux de pages qualifiées varient entre 0 et 1. Ces courbes montrent une corrélation positive entre la performance et le taux de pages qualifiées, surtout pour les valeurs de performance importantes (supérieures à 0,6).
- pour la stratégie du lien complet, les taux de pages qualifiées varient entre 0 et 0,4, ce qui est extrêmement faible. Aucune corrélation n'apparaît sur ce graphique entre la performance et le taux de pages qualifiées.

Pour les points remarquables des graphiques de la figure 5.4 (points entourés), les taux de pages qualifiées atteignent 0,35 pour le simple lien et 0,56 pour le lien moyen, ce qui est très faible.

5.3.4 Discussion et limites

L'évaluation menée sur ce corpus montre que la méthode 1 permet d'obtenir une propagation de bonne qualité. Cependant les valeurs de performance

ne sont pas satisfaisantes. En effet, cette méthode propose d'intervenir manuellement au moins deux fois dans chaque agrégat du seuil de départ K, et plus si les valeurs partagées par les éléments les plus distants sont différentes. Rappelons que les méthodes de classification appliquées à ce corpus forment des agrégats homogènes de faible taille, parfois composés de 3 ou 4 pages seulement. Ceci entraîne un trop grand nombre de pages indexées manuellement par rapport au nombre de pages indexées par propagation. Une des façons d'améliorer la performance consiste à augmenter le paramètre α, nombre minimum de pages que l'agrégat doit contenir pour appliquer la méthode. Pour α fixé à 5, les résultats sont sensiblement meilleurs pour la performance. Les meilleurs résultats observés pour la stratégie du lien moyen sont présentés sur la figure 5.6 : pour une qualité supérieure à 0,8, on aperçoit une série de points proches de 0,8. Par contre, le taux de pages qualifiées reste toujours faible : la valeur 0,51 est observée dans le meilleur des cas ; les autres points se situent entre 0,4 et 0,5.

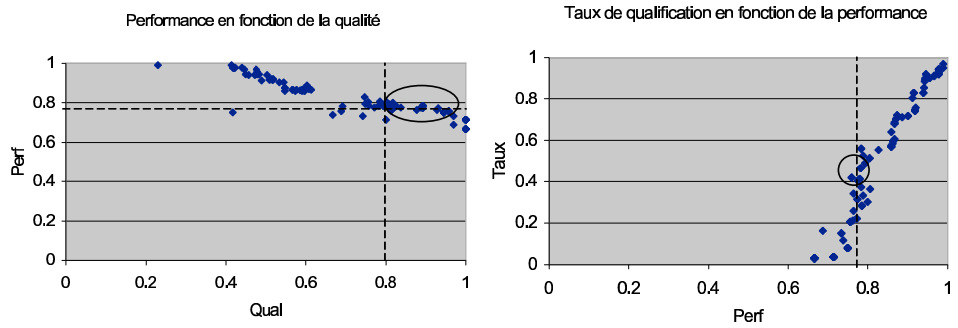

FIG. 5.6 – Résultats pour la stratégie du lien moyen en posant $\alpha = 5$ (méthode 1)

De plus, la propagation n'a lieu que si les pages les plus éloignées partagent les mêmes valeurs de métadonnées. Cette condition relativement forte, permet de limiter le nombre d'erreurs lors de la propagation, mais force la propagation des valeurs à des niveaux inférieurs dans le dendrogramme. Ceci limite le nombre de pages concernées par la qualification et explique pourquoi le taux de qualification est si faible.

5.4 Méthode 2

5.4.1 Présentation de la méthode

La première méthode offre l'avantage d'utiliser toute la richesse de la hiérarchie des dendrogrammes et ne se limite pas à l'étude d'un seul niveau de coupure. Cependant, les résultats obtenus pour la performance et le taux de

pages qualifiées ne sont pas satisfaisants. La seconde méthode propose d'améliorer ces indices en intervenant le moins possible manuellement et en propageant à des niveaux plus élevés.

Cette seconde méthode opère pour un niveau t donné, auquel correspond une partition π_t d'agrégats. Elle repose sur l'hypothèse que l'élément central de chaque agrégat, c'est-à-dire l'élément le plus proche de tous les autres, est le plus représentatif des éléments de l'agrégat. Dans cette méthode, ce sont les valeurs de métadonnées de l'élément central qui sont propagées. La *centralité de proximité* d'une page P_i se calcule par la fonction introduite par Sabidussi (équation 2.3, page 21) :

$$C_p(P_i) = \frac{n-1}{\sum_{j=1}^{n} d(P_i, P_j)} .$$

5.4.2 Algorithme de la méthode

L'algorithme que nous proposons pour cette seconde méthode est le suivant.

Soit π_t la partition obtenue au niveau t,

Pour chaque agrégat de π_t composé de n pages, noté $Ag = \{P_i / 1 \leq i \leq n\}$ avec $n \geq 2$

1. **Pour** chaque page P_i

   Calcul de sa valeur de centralité :
   $C_p(P_i) = \frac{n-1}{\sum_{j=1}^{n} d(P_i,P_j)}$

   Fin Pour

   Soit $E = \{C_p(P_1); ...; C_p(P_i); ...; C_p(P_n)\}$ l'ensemble des valeurs de centralité prises par les pages de l'agrégat.

2. indexation manuelle pour les métadonnées choisies de la page (ou d'une des pages) ayant pour valeur de centralité $max(E)$;
3. propagation des valeurs aux autres pages de l'agrégat.

Fin Pour

5.4.3 Evaluation

Pour les raisons évoquées précédemment et pour permettre une comparaison entre les deux méthodes, l'évaluation de la seconde méthode est réalisée sur le même corpus.

5.4.3.1 Propagation

Nous procédons à la propagation des valeurs métadonnées conformément à l'algorithme décrit dans la section 5.4.2. La propagation des métadonnées a été réalisée pour les 3 dendrogrammes, ainsi que pour tous les niveaux t possibles. Les métadonnées choisies sont les mêmes que pour l'évaluation de la première méthode : *Type d'autorité*, *Type d'information* et *Type de site*. Comme précédemment, à chaque fois que la méthode doit faire intervenir une indexation humaine, nous utilisons les résultats de l'indexation réalisée dans la section 4.4 (page 78). Si au sein d'un agrégat, la page centrale contient une valeur indéterminée, nous utilisons alors la valeur (de la métadonnée concernée) de la page suivante dans le classement des pages en fonction de leur centralité décroissante.

Notre implémentation contient aussi deux modules :
- un module principal qui exécute l'algorithme de propagation (5.4.2) : il donne le nombre de valeurs de métadonnée propagées et le nombre de valeurs indexées manuellement ;
- un module vérifiant les résultats de la propagation : il donne le nombre de valeurs de métadonnées propagées identiques à celles obtenues par la qualification manuelle (valeurs justes), le nombre de valeurs différentes (valeurs fausses) et le nombre de valeurs pour lesquelles nous ne pouvons pas conclure (cas où la valeur indexée manuellement est indéterminée).

Ainsi la méthode de propagation répartit les $3N$ valeurs de métadonnées en cinq cases (cf. Tab. 5.1, page 109). Précisons que :
- la valeur v_j^{non-p} (nombre de valeurs de métadonnées qualifiées manuellement) est supérieure ou égale à 3 fois le nombre d'agrégats de π_t ; en effet, un élément central peut comporter une valeur indéterminée ;
- la valeur v_f^{non-p} (nombre de valeurs de métadonnées non qualifiées) est prévisible en examinant le dendrogramme : elle est égale à trois fois le nombre de singletons au niveau t.

5.4.3.2 Présentation des résultats

Pour évaluer notre méthode, nous utilisons les indices présentés dans la section 5.3.3.2. Les résultats de cette expérimentation sont présentés sur les figures 5.7, 5.8 et 5.9.

La figure 5.7 donne les valeurs de qualité obtenues pour chaque niveau t décroisant ($1 \leq t \leq T$) pour les trois stratégies de classification :
- les meilleurs résultats sont obtenus pour la stratégie du lien complet. Ils varient entre 0,87 et 1. Nous observons que la qualité croît régulièrement au fur et à mesure que le niveau t décroît. Pour les quatre niveaux inférieurs, la méthode propage les informations sans erreur.

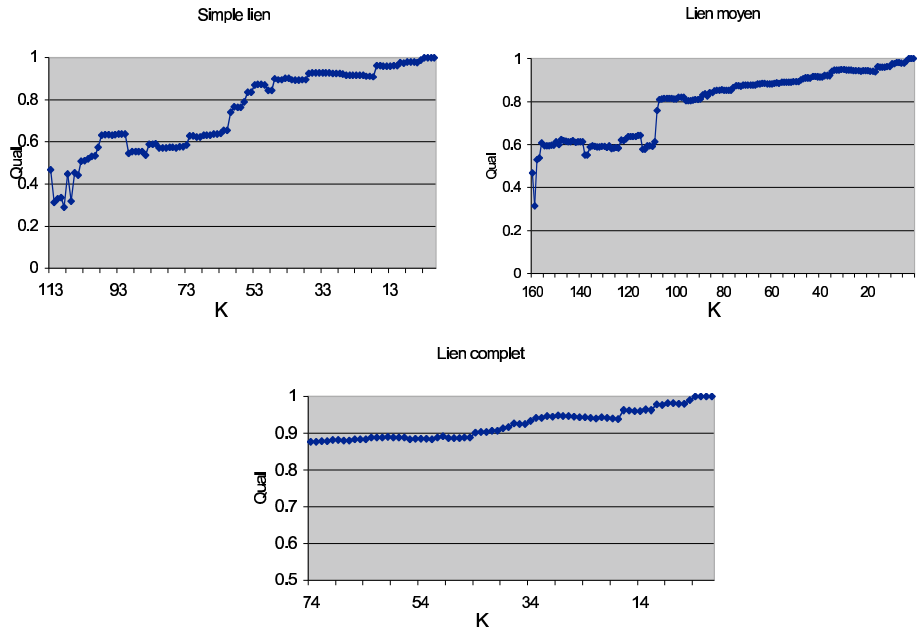

FIG. 5.7 – Représentation de la qualité pour les trois statégies (méthode 2)

- pour la stratégie du lien moyen, les résultats varient entre 0,31 et 1. Pour les quatre niveaux supérieurs, les valeurs de qualité sont très basses. L'évolution de la qualité se fait ensuite de manière irrégulière où les résultats oscillent autour de 0,6. A partir du niveau 107, c'est-à-dire pour 66% des niveaux t, la méthode propage les informations avec moins de 20% d'erreurs. Pour les quatre niveaux inférieurs, la propagation s'effectue sans erreur.
- pour la stratégie du simple lien, les résultats varient eux aussi entre 0,31 et 1. Pour les dix niveaux supérieurs, les valeurs de qualité sont irrégulières et très basses (inférieures à 0,5). L'évolution de la qualité se fait toujours de manière irrégulière avec des palliers oscillant autour de 0,6. A partir du niveau 60, la qualité croît plus régulièrement et à partir du niveau 55, c'est-à-dire pour 48% des niveaux t, la méthode propage les informations avec moins de 20% d'erreurs. Pour les quatre niveaux inférieurs, la propagation s'effectue sans erreur.

Nous remarquons que pour la méthode simple lien et pour celle du lien moyen, les résultats de la qualité aux niveaux supérieurs ne sont pas réguliers. En effet, à ces niveaux les agrégats sont de taille importante. Cependant, ils ne sont pas encore stables et leur cohésion est mauvaise. Leurs éléments centraux ne représentent pas systématiquement la majorité des pages, et c'est pourquoi, d'un seuil à l'autre la qualité oscille.

La figure 5.8 donne l'évolution de la performance $Perf$ en fonction de la

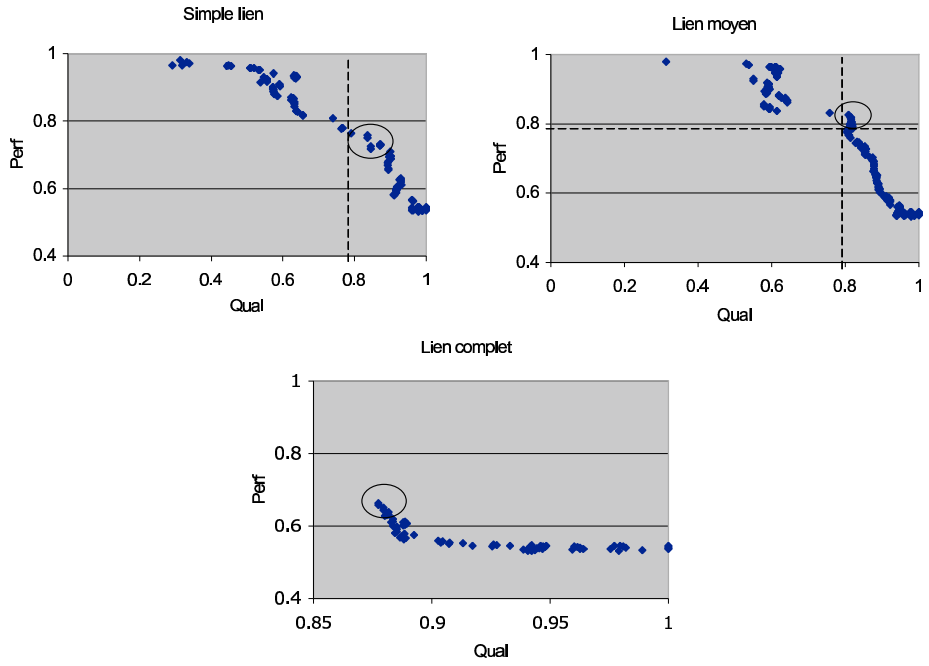

FIG. 5.8 – Représentation de la performance en fonction de la qualité pour les 3 stratégies (méthode 2)

qualité *Qual* pour les trois stratégies.

- Pour les stratégies du simple lien et du lien moyen, les courbes présentent une forme similaire. Les valeurs de performance varient entre 0,55 et 1. Ces graphiques montrent dans la partie supérieure gauche des groupes de points partageant une mauvaise qualité ($0,31 \leq Qual \leq 0,77$) et une bonne performance. Il s'agit des résultats obtenus pour les niveaux supérieurs des dendrogrammes. Ces seuils ont un faible nombre d'agrégats, peu de pages sont indexées manuellement et c'est pourquoi leur performance est élevée. Par contre, leur cohésion est mauvaise, ce qui entraîne une mauvaise qualité de propagation. A partir d'une qualité proche de 0,8, une liaison fonctionnelle entre la qualité et la performance apparaît (surtout pour la méthode du lien moyen). La performance décroît au fur et à mesure que la qualité croît. Cette corrélation négative s'explique : à partir d'une qualité de 0,8, les agrégats sont de plus en plus stables mais leur taille décroît progressivement, ce qui explique la baisse de performance.
- pour la stratégie du lien complet, les valeurs de performance varient entre 0,53 et 0,66, ce qui est faible. Une corrélation négative apparaît pour les valeurs de performance comprises entre 0,66 et 0,56 et s'explique comme précédemment. Ensuite la performance converge vers la valeur minimale possible pour cette seconde méthode : 0,5 (une page qualifiée manuellement, une page qualifiée par propagation).

Ces courbes montrent que pour une qualité convenable (supérieure à 0,80), les meilleurs résultats de la performance[4] atteignent 0,82 pour le lien moyen (0,81 ; 0,82), 0,76 pour le simple lien (0,84 ; 0,76) et 0,67 pour le lien complet (0,88 ; 0,67). Les résultats les plus intéressants concernent la méthode du lien moyen, pour laquelle il existe sept points remarquables où la qualité et la performance sont toutes les deux supérieures à 0,8 ($0,809 \leq Qual \leq 0,817$ et $0,8 \leq Perf \leq 0,825$).

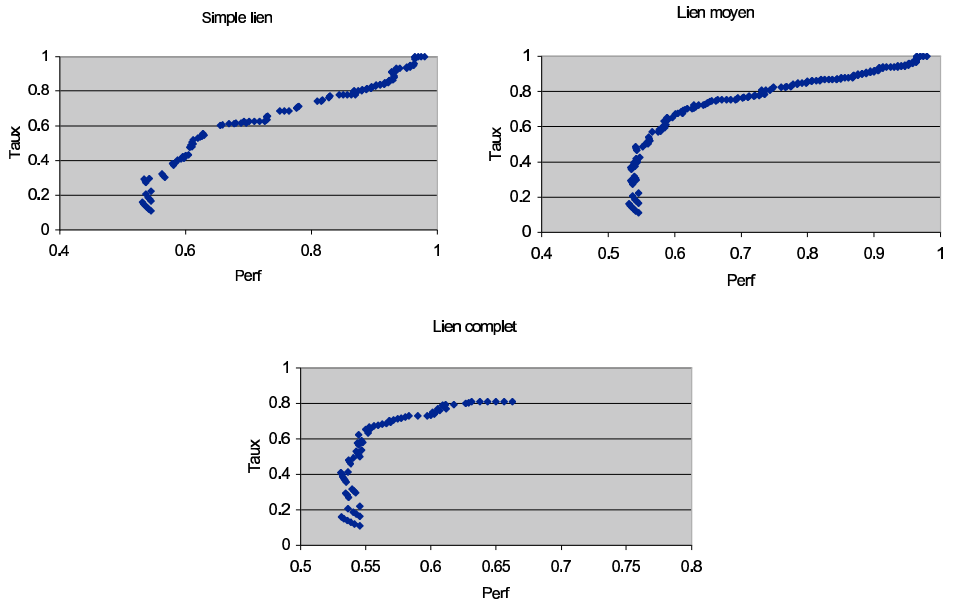

FIG. 5.9 – Représentation du taux de qualification en fonction de la performance pour les 3 stratégies

La figure 5.9 montre l'évolution du taux de pages qualifiées en fonction de la performance. Pour les trois stratégies les courbes sont croissantes. Elles varient entre 0,11 et 1 pour le lien moyen et le simple lien, et entre 0,11 et 0,88 pour le lien complet. Ces courbes montrent une corrélation positive entre la performance et le taux de pages qualifiées. Pour les points entourés des graphiques de la figure 5.8, le taux de pages qualifiées atteint la valeur maximum de 0,68 pour le simple lien, la valeur maximum de 0,85 pour le lien moyen, et la valeur maximum de 0,80 pour le lien complet. De plus, pour les sept points remarquables observés pour la stratégie du lien moyen, le taux de performance est lui aussi supérieur à 0,8.

[4]Les meilleurs résultats de performance pour une qualité supérieure à 0,8 sont entourés sur les graphiques du simple lien, du lien moyen et du lien complet (figure 5.8)

5.5 Comparaison des méthodes

L'évaluation de nos deux méthodes de propagation sur le corpus d'astronomie montre qu'il est possible de propager des valeurs de métadonnées avec une bonne qualité. Pour les deux méthodes, les résultats de la qualité sont conditionnés par l'exigence de la stratégie d'agrégation. Les meilleurs résultats sont observés pour la stratégie du lien complet.

Cependant, ces expériences montrent qu'il est impossible d'avoir à la fois une très bonne qualité et une bonne performance. Pour obtenir une bonne performance il faut introduire du bruit, c'est-à-dire des erreurs dans la propagation. La figure 5.10 donne les résultats obtenus par les deux méthodes pour le rapport qualité/performance. Sur ces graphiques, les résultats de la première méthode sont matérialisés par des cercles, tandis que ceux obtenus par la seconde sont matérialisés par des croix. Cette figure montre que la seconde méthode donne de meilleurs résultats que la première à la fois pour la majorité des seuils et pour les trois stratégies. A qualité égale, la seconde méthode est donc plus performante que la première.

La figure 5.11 permet de comparer les résultats obtenus pour le rapport taux de qualification/performance. Comme précédemment, les résultats de la première méthode sont matérialisés par des cercles, tandis que ceux obtenus par la seconde sont matérialisés par des croix. Ces graphiques indiquent que pour la majorité des seuils, les résultats sont nettement meilleurs pour la seconde méthode. A performance égale, le taux de pages qualifiés est plus important pour la seconde méthode que pour la première.

5.6 Pour conclure

Dans ce chapitre, nous nous sommes intéressés à la qualification semi-automatique de pages web.
D'un point de vue théorique,

- nous avons présenté deux méthodes de propagation de métadonnées basées sur la structure du graphe du Web. Ces méthodes s'effectuent après structuration du corpus par le principe de *co-sitation* et utilise la notion de distance dans les graphes ;
- nous avons aussi proposé trois indices permettant l'évaluation de ces méthodes. Ces indices reflètent :
  * la qualité de la propagation qui mesure si la qualification automatique donne de bons résultats ;
  * la performance qui quantifie l'intervention humaine dans la méthode. La méthode est d'autant plus performante, que l'humain intervient

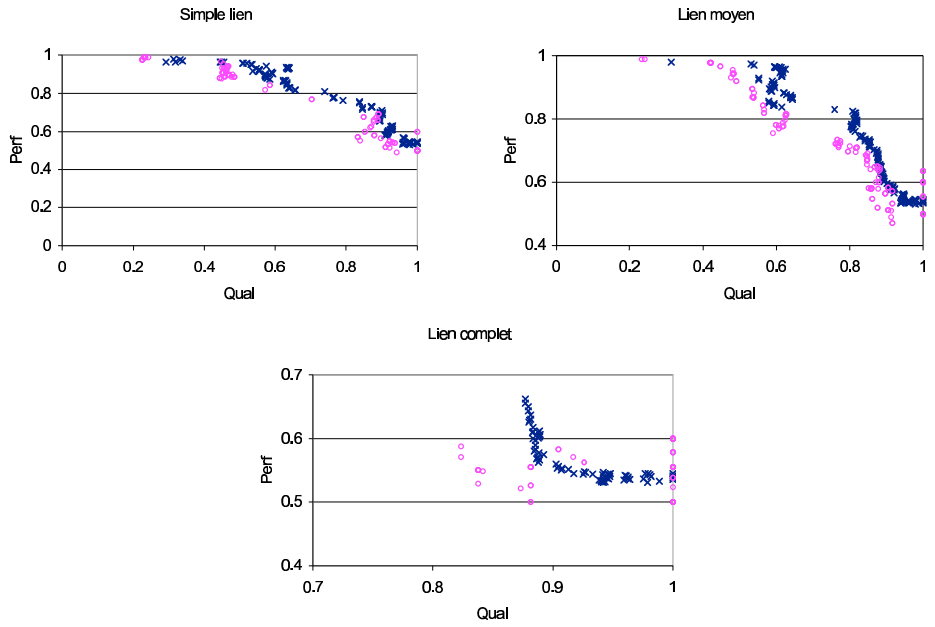

FIG. 5.10 – Performance en fonction de la qualité : résultats comparatifs des deux méthodes de propagation

peu ;

* le taux de pages qualifiées qui mesure le nombre de pages qualifiées par rapport au nombre de pages du corpus.

D'un point de vue pratique, nous avons implémenté ces méthodes et procédé à leur évaluation sur le corpus formé au chapitre 4. Les résultats obtenus pour la seconde méthode et pour la stratégie du lien moyen sont particulièrement encourageants. En effet, nous avons observé plusieurs seuils pour lesquels la qualité, la performance et le taux de qualification sont tous les trois supérieurs ou égaux à 80%. Si ces résultats se confirmaient sur d'autres corpus, le taux de qualification qui est prévisible[5], pourrait devenir un critère de coupure[6] de la clusterisation. Nous pensons qu'il serait intéressant de poursuivre dans cette voie et de tester la seconde méthode sur d'autres corpus et surtout pour d'autres métadonnées comme le thème par exemple.

[5]Dans la seconde méthode, le taux de qualification est directement lié à la clusterisation. En effet, la valeur v_f^{non-p} est égale à trois fois le nombre de singletons au niveau t.

[6]*cut-off* en anglais.

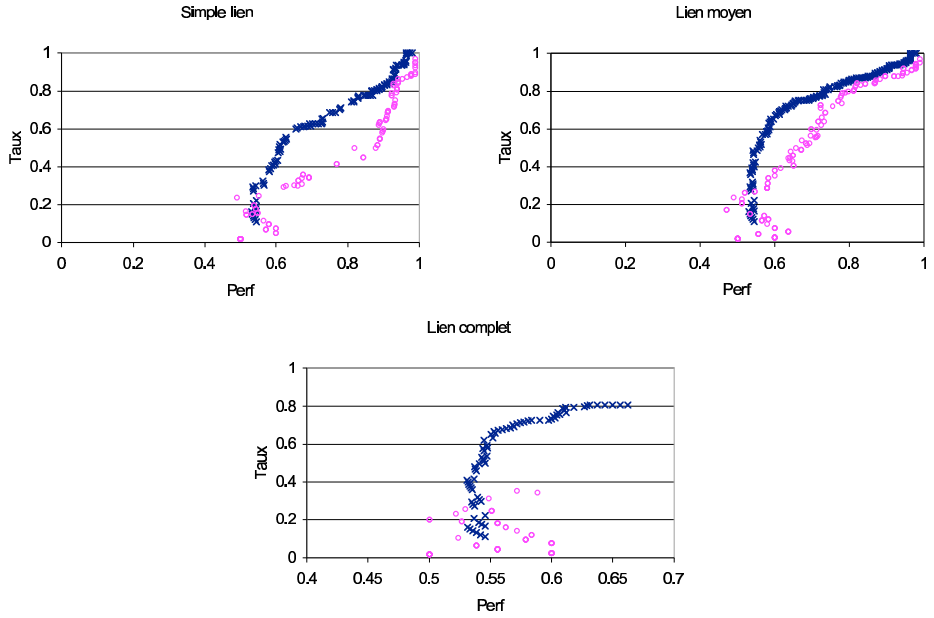

FIG. 5.11 – Taux de qualification en fonction de la performance : résultats comparatifs des deux méthodes de propagation

Chapitre 6

Conclusion

Ce chapitre de conclusion rappelle le contexte de notre travail – la recherche d'information sur la Toile – et les difficultés soulevées (section 6.1). Il résume ensuite nos contributions (section 6.2) et présente les limites de notre travail d'un point de vue théorique et expérimental (section 6.3). Nous pensons que nous ne pouvons pas clore cette thèse sans discuter de la possibilité d'un passage à l'échelle de notre approche. La dernière section de ce chapitre (section 6.4) est une discussion sur ce sujet, menée à partir des premiers résultats d'une étude exploratoire effectuée à très grande échelle (sur un corpus de plus de 5 millions de pages web).

6.1 Problème abordé

Nul ne peut nier les avantages apportés par le réseau Internet à la société de l'information. La rapidité des échanges de données, la gratuité, la couverture mondiale et la diversité des services offerts (messagerie, web, etc.) sont autant d'éléments qui ont contribué à son succès. Plus particulièrement, le Web ou Toile d'araignée mondiale, est devenue en quelques années une source d'information incontournable. Devant l'accroissement du volume d'information disponible, l'Internet a besoin plus que jamais d'outils de recherche d'information performants pour satisfaire ses utilisateurs. Dans un premier temps, les techniques « classiques » de recherche d'information, basées sur le contenu lexical des documents ont été transposées au Web. Les moteurs de recherche dits de « première génération », comme Altavista en 1995, voient le jour. Ce type de moteur contient des index qui comportent à la fois la liste des termes retrouvés sur les pages et leurs places dans celles-ci, permettant ainsi l'utilisation des opérateurs booléens et de proximité. Dans un second temps, une nouvelle génération de moteurs apparaît. Leur particularité commune est de ne pas se limiter au contenu lexical des ressources, mais d'exploiter les autres données issues du Web et de son utilisation. Ces données sont les informations concernant l'usage (fichiers de log, cookies, etc.) mais surtout les informations portées par la

structure du Web. En effet, le Web se présente comme un gigantesque système hypertexte dans lequel les pages web sont reliées les unes aux autres par l'intermédiaire de liens. Formellement, il peut être représenté par un graphe orienté. L'application la plus connue de l'analyse des liens du Web pour la recherche d'information est celle qui concerne le classement des pages. Implémenté dans le moteur de recherche Google, cet algorithme de classement de pages est une petite révolution pour la recherche d'information sur la Toile. Selon ses concepteurs, il traduit la popularité des pages web. Les pages recevant de nombreux liens hypertextes, c'est-à-dire les pages très citées, voient leur score de pertinence augmenté.

Quel que soit le fonctionnement des outils de recherche disponibles sur la Toile, aucun de ceux-ci ne semble véritablement prendre en compte le caractère hétérogène des ressources disponibles. Tous s'appuient, comme les Systèmes de Recherche d'information traditionnels (SRI), sur une représentation sémantique des documents. Contrairement aux bases documentaires traditionnelles, le Web se présente comme un gisement d'information non contrôlé et sans aucune gestion. Ainsi les ressources retrouvées sont hétérogènes à tout point de vue, au niveau de leur contenu thématique bien sûr, mais aussi au niveau de leur genre, de leur langue, de leur niveau, du public visé, etc. Les utilisateurs ont recours aux moteurs de recherche avec des attentes et des objectifs bien différents, et ne sont pas toujours satisfaits des ensembles de résultat retournés par les moteurs.

Dans cette thèse, nous nous sommes intéressés aux difficultés de recherche d'information engendrées par le caractère hétérogène de ce nouveau médium. Notre démarche vise une description plus complète des ressources au delà de leur description sémantique. Notre objectif final étant une qualification systématique des ressources par l'affectation de métadonnées non thématiques. Dans le cadre de cette thèse, nous avons considéré deux niveaux pour la caractérisation des ressources : la page web, d'une part, car elle se présente comme une unité informationnelle autonome et le site web, d'autre part, car il est selon nous le regroupement de pages le plus évident sur la Toile.

6.2 Contributions

Nos contributions à travers cette thèse peuvent se résumer en quatre points :

1. L'étude des parallèles et des analogies possibles entre les champs disciplinaires de l'analyse des réseaux sociaux, de la bibliométrie citationiste et de l'étude du graphe du Web. Parmi les rapprochements effectués entre ces disciplines, l'analogie entre le graphe de citation représentant le vaste réseau des publications scientifiques et le graphe du Web semble la plus prometteuse. Elle ouvre de nombreuses perspectives pour la recherche d'information sur la Toile.

2. La présentation d'une typologie possible pour décrire les pages web. Elle comporte quatre métadonnées qui sont relatives aux pages et plus largement aux sites web dont elles font partie. Ces métadonnées sont les suivantes : le type de site dont fait partie la page (qui traduit le rôle informationnel du site), le type d'autorité responsable du site, le type d'information contenue dans la page et le type de page lié à ses caractéristiques physiques. Pour chacune de ces métadonnées, un ensemble de valeurs possibles est proposé.

3. La proposition d'une approche de caractérisation collective des pages web comprenant deux étapes. La première étape, l'extraction de corpus homogènes, vise à rapprocher des pages partageant des caractéristiques communes. La seconde étape, l'affectation semi-automatique de métadonnées au sein de chaque corpus homogène, est basée sur la propagation : au départ, seule une faible proportion des ressources sont qualifiées, leurs informations sont ensuite propagées aux autres ressources du corpus.
   Au niveau méthodologique, l'extraction des corpus homogènes est effectuée par l'application de méthodes de classification hierarchique ascendante provenant de l'analyse de données. Le point fondamental de notre approche est l'utilisation d'une similarité basée sur la structure hypertexte du Web, plus particulièrement basée sur le principe de *co-sitation*. Ce principe est la transposition sur le Web de la méthode des co-citations bien connue en scientométrie. Il repose sur l'hypothèse d'une auto-organisation de la Toile analogue à l'hypothèse de la science combinatoire évoquée par Price. Le Web, comme l'univers des publications scientifiques, est un espace multi-auteurs dans lequel les ressources peuvent se positionner les unes par rapport aux autres grâce aux liens hypertextes. Selon ce principe deux pages sont proches et suceptibles de partager des propriétés communes, si par rapport à leurs fréquences de citations respectives, leur fréquence de co-citation est importante.
   Deux méthodes d'affectation semi-automatique de métadonnées au sein des corpus homogènes sont investies. Toutes les deux s'appuient sur la structure des graphes de co-citation induits par chaque corpus homogène. Elles différent par le choix des pages à indexer manuellement et par la façon dont sont propagées les métadonnées. Pour chacune de ces méthodes, la notion de distance dans les graphes de co-citation est un critère pour le choix des pages à indexer manuellement.

4. L'évaluation de notre approche. Cette évaluation a été réalisée sur un corpus provenant du Web et construit à partir du moteur de recherche Google. Les différentes expérimentations menées sur ce corpus montrent l'intérêt de notre approche pour la caractérisation des pages web. Premièrement, la construction de corpus homogènes par le principe de *co-sitation* s'est avérée possible pour trois des quatre métadonnées proposées dans la typologie. Deuxièmement, les tests des méthodes de propagation ont donné des résultats encourageants, en particulier pour la seconde méthode. Pour une qualité de propagation convenable et équivalente à la première méthode, celle-ci s'est avérée moins coûteuse pour l'intervention humaine et tend à

qualifier un plus grand nombre de pages du corpus.

6.3 Limitations

Au niveau théorique, les limitations de notre approche sont directement liées à celles de l'analyse des *sitations*. Une des limites importantes de l'analyse des *sitations* concerne les *sitations* vides de sens, c'est-à-dire celles qui sont formées sans motivation de communication particulière (liens gratuits, liens de publicité). De l'utilisation d'un indice basé sur la co-citation découle un intérêt majeur : si les liens vides de sens sont nombreux sur le web, les *co-sitations* de pages formées sans motivation particulière, ne peuvent pas semble-t-il avoir de fréquences importantes.

Les autres limites de l'analyse des *sitations* entraînent des conséquences plus importantes pour notre approche. Ces limites sont : premièrement, le fait que seuls les points d'entrée sont majoritairement cités ; deuxièmement, le fait que la probabilité pour qu'un point d'entrée ne soit jamais co-cité est non négligeable. De ces deux limites proviennent les restrictions de corpus que nous avons constatées lors de nos expérimentations. Notre approche ne peut pas prétendre à la caractérisation de l'ensemble des pages du Web, mais seulement de celles apparaissant comme intéressantes à citer, en général celles correspondant à des points d'entrées sur les sites. Il existe toutefois des avantages à la caractérisation des points d'entrées seulement. D'une part, les points d'entrée sont généralement accessibles par des URLs stables et qui perdurent dans le temps. D'autre part, même si l'information contenue sur les points d'entrée est mise à jour, leurs caractéristiques typologiques sont rarement remises en cause. Par exemple, un site homeserveur se voit difficilement devenir un site de recherche, etc.
Au niveau expérimental, la limite la plus évidente concerne la manière dont a été formé le corpus de test et l'impossibilité d'obtenir un sous-graphe exhaustif du Web avec les moteurs de recherche disponibles. Plus globalement, les autres difficultés concernent l'obtention d'un graphe du Web unique et « propre » (c'est-à-dire sans pages doublons, sans ULRs alias, etc.) et l'impossibilité d'identifier automatiquement les citations faites entre les sites web sans commettre d'erreur.

6.4 Perspectives : vers un passage à l'échelle

Pour répondre aux difficultés de recherche d'information liées au caractère hétérogène du Web, nous avons proposé une démarche semi-automatique de caractérisation des sites et des pages web. Cette démarche se base sur l'analyse des liens, c'est-à-dire sur les relations existantes entres les pages web. Les différents tests de la méthode présentée (extraction de corpus homogènes et propagation) ont été réalisés sur un corpus de petite taille, composé de 198 points

d'entrée seulement. Les résultats encourageants permettent d'envisager un passage à l'échelle et de réfléchir à l'intégration de notre méthode par des outils de recherche. Pour évaluer, d'une part, les limites de l'analyse des *sitations* évoquées au chapitre 3 (section 3.2.3, page 61) sur un corpus représentatif du Web, et envisager d'autre part, la perspective d'un passage à l'échelle, nous avons étudié les caractéristiques d'un graphe de citation de très grande taille.

6.4.1 La collection

Pour cette étude exploratoire, nous avons utilisé un corpus nommé Wfr4 et composé de 5.057.642 pages. Ces pages ont été collectées sur la Toile en décembre 2000 grâce à un robot[1] développé par M. Mathias Géry et M. Dominique Vaufreydaz, membres du laboratoire CLIPS[2] de l'université de Grenoble. Toutes ces pages font partie de domaines d'origine géographique francophone (Tab. 6.1), ce qui ne signifie pas que tous les documents sont en langue française. En effet, de nombreux sites proposent plusieurs versions de leurs documents dans plusieurs langues.

Extensions	.fr	.be	.lu
Pays	France	Belgique	Luxembourg
Nombre de sites	29.441	8.851	1.152

TAB. 6.1 – Extensions les plus représentées dans la collection

6.4.2 La découverte du graphe

La découverte du graphe, c'est-à-dire des relations entre les pages de cette collection, a été réalisée au sein de notre laboratoire. Les différentes étapes ont été les suivantes.

- L'extraction des URLs des pages citantes[3] et des pages citées.
- La normalisation des URLs, comme par exemple l'écriture en minuscule des noms de machine ou la suppression du numéro de port lorsqu'il s'agit du port par défaut, etc. Pour réduire la taille du corpus (nombre de sommets du graphe de citation), les URLs contenant des requêtes sont tronquées au point d'interrogation. Le terme que nous employons pour désigner ces URLs tronquées est *netpath*[4]. Les URLs tronquées

[1] `http://www-mrim.imag.fr/membres/mathias.gery/CLIPS-Index/`

[2] `http://www-clips.imag.fr/`

[3] Les pages retrouvées par le robot mis au point par M. Géry et D. Vaufreydaz sont stockées dans 5060 fichiers au format texte. Sont reportés dans ces fichiers pour chaque page : des **méta-informations**, comme par exemple, l'URL où a été retrouvée la page (URL citante), et le **code source** de la page (code HTML).

[4] Les spécifications des URL [rfc, 2004] indiquent que la forme la plus générale est composée de neuf champs : *<scheme>*://*<user>*:*<passwd>*@*<host>*:*<port>*/*<path>*;*<params>*?*<query>*#*<fragm>*. Pour

donnent les chemins pour accéder aux ressources, mais ne donne pas les requêtes et les arguments possibles. Dans la majorité des cas, les URLs ainsi obtenues correspondent encore à des pages web. Cette opération réduit la taille du corpus de 5.057.642 pages à 3.823.589 netpaths.
- La construction de tables associatives qui associent un numéro à chaque netpath et à chaque site. Dans cette expérience, un site est défini comme la concaténation du nom de machine et du nom de domaine (*<host>*).

Outre les difficultés évoquées au chapitre 3 (section 3.2.3) comme par exemple les erreurs d'écriture dans les pages, l'acquisition du graphe n'a pas posé de difficulté technique insurmontable. L'obtention de ce graphe, avec une machine de capacité moyenne (512 méga-octets de mémoire centrale, 1 GHz de vitesse de processeur) a été réalisée en 40 heures environ. Les nœuds de ce graphe sont des netpaths et les arcs les relations externes entre ces netpaths.

6.4.3 Les caractéristiques du graphe de *sitation*

Nous rappelons dans un premier temps les caractéristiques de la collection Wfr4. Elle contient :
- 5.057.642 pages web,
- 3.823.589 netpaths,
- 43.462 sites web (nom de domaine).

La figure 6.1 montre les distributions du nombre de pages et du nombre de netpaths par site. Ces distributions sont conformes aux lois de l'information (lois hyperboliques). Nous nous étonnons cependant devant la quantité de sites contenant un très faible nombre de pages : 20.679 sites ne contiennent qu'une seule page (soit 47,6% des sites de la collection), et 20.748 sites ne contiennent qu'un seul netpath (soit 47,7% des sites de la collection).

Le graphe de netpaths est un graphe orienté (non valué) qui contient :
- 1.004.152 sommets (netpaths),
- 3.324.703 arcs (relations externes entre les netpaths),
- 831.009 sommets citants, c'est-à-dire des sommets pour lesquels $d_s > 0$; les sommets citants appartiennent à 18.834 sites,
- 250.558 sommets cités (points d'entrée), c'est-à-dire des sommets pour lesquels $d_e > 0$; les sommets cités appartiennent à 43.312 sites[5].

ce qui nous concerne, nous ne nous intéressons qu'à des pages, donc le *<scheme>* est toujours `http`. Les pages ont été collectées par un robot, donc les champs *<user>* et *<passwd>* sont toujours vides. Enfin, nous ne nous intéressons qu'aux pages effectives, donc nous n'utilisons pas le champ *<query>* qui ne concerne que les pages dynamiques, ni le champ *<fragm>* qui permet de spécifier une sous-partie contiguë dans une page (statique ou dynamique). Ce qui fait que pour nous, il reste les champs : *<host>*:*<port>*/*<path>* que nous appelerons globalement *netpath*.

[5]Rappelons que le corpus a été formé par le parcours d'un robot qui a suivi les liens. Tous les sites présents dans le corpus ont été atteints par le robot et devraient avoir un point d'entrée. La présence de 150 sites apparaissant sans point d'entrée s'explique peut-être par le phénomène des noms de domaine alias.

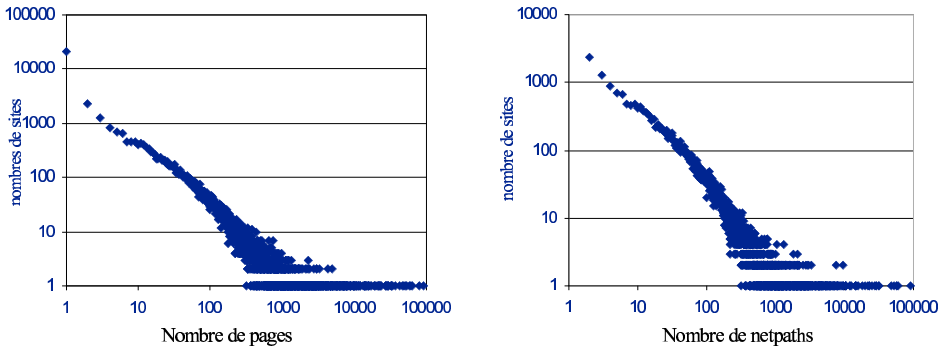

FIG. 6.1 – Distribution du nombre de pages et du nombre de netpaths par site

Nous remarquons que sur les 3,82 millions de netpaths de la collection Wfr4, seuls 1 million environ entretiennent des relations avec d'autres sites du corpus (reçoivent ou émettent des *sitations*). De plus, parmi ces 3,82 millions de netpahs, seulement 250.558 netpaths reçoivent des citations externes et se présentent ainsi comme des points d'entrée. Ce chiffre nous permet d'avancer une estimation de 6,5% pour la proportion de pages (netpaths) points d'entrée sur le web, c'est-à-dire la proportion maximale de pages pouvant être qualifiées par notre approche.

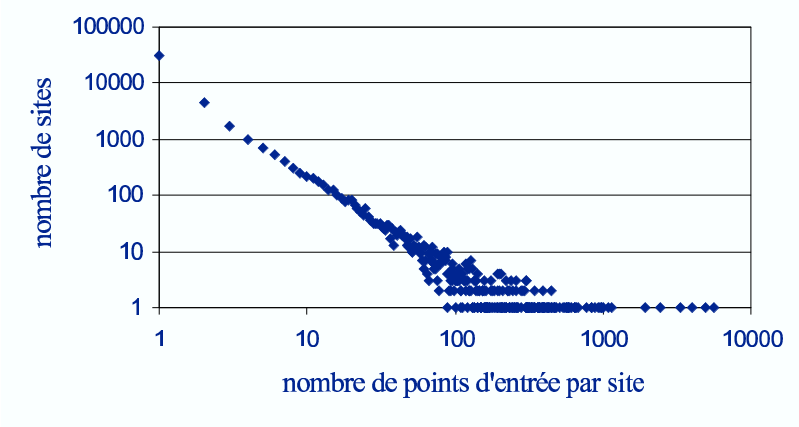

FIG. 6.2 – Distribution du nombre de points d'entrée par sites

La figure 6.2 donne la distribution du nombre de points d'entrée par site. Cette distribution est aussi conforme aux lois de l'information. Nous remarquons sur la figure 6.3 qu'il n'y a aucune corrélation entre le nombre de total de netpaths par site et le nombre total de points d'entrée par site (coefficient de corrélation égal à 0,46). De manière générale, nous observons que lorsqu'un site possède plusieurs points d'entrée, seuls quelques uns ont une fréquence de

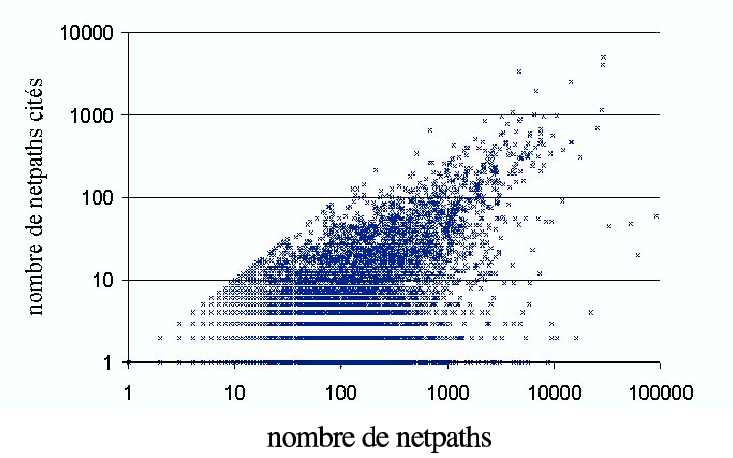

FIG. 6.3 – Nombre de points d'entrée en fonction du nombre de netpaths par site

citation très élevée. La figure 6.4 donne pour illustrer ce propos la distribution du nombre de citations reçues par les points d'entrée de deux sites, composés respectivement de 113 et de 789 netpaths, et comportant respectivement 17 et 30 points d'entrée.

La figure 6.5 montre pour les netpaths les distributions des degrés entrants et sortants. Bien que l'on ne prenne en compte que les citations externes aux sites (*les sitations*), les ajustements obtenus se rapprochent des résultats présentés par Border et al. [Broder et al., 2000] et Albert et al. [Albert et al., 1999] (section 2.3.1.2, page 43).

6.4.4 Vers la construction du graphe de *co-sitation*

Une condition nécessaire pour qu'une page engendre des *co-sitations* est qu'elle doit émettre au moins deux *sitations*. Ainsi toutes les pages ne satisfaisant pas cette condition sont éliminées. Dans la collection Wfr4, 384.655 netpaths n'émettent qu'un seul lien externe. Dans le graphe de *sitations* 384.655 relations de *sitations* (arcs) sont donc supprimées, entraînant la disparition de 384.655 sommets citants (soit 46% des sommets citants) et 22.306 sommets cités, c'est-à-dire de 8,9% des points d'entrée. Le nouveau graphe de *sitation* ainsi obtenu contient :

- 635.535 sommets (netpaths),
- 2.940.048 arcs (relations entre les sommets),
- 446.354 sommets citants, pour lesquels $d_s > 1$; ces sommets appartiennent à 12.229 sites.
- 228.252 sommets cités et co-cités ; ces sommets appartiennent à 40.239 sites.

Ces indications nous permettent de conclure que :

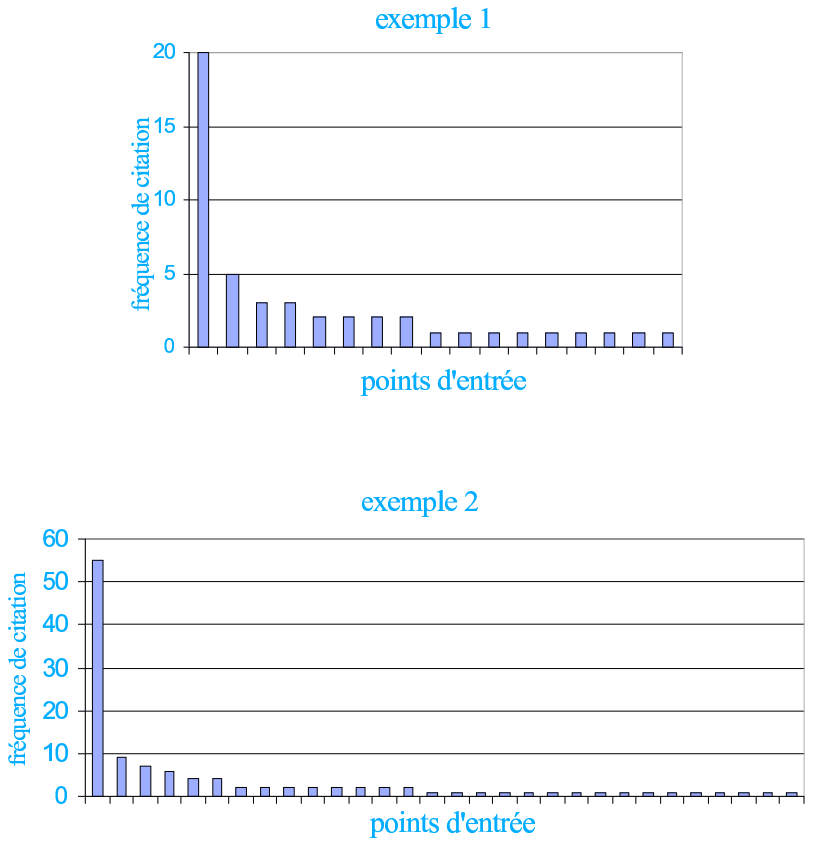

FIG. 6.4 – Distribution des fréquences de citations reçues par les points d'entrée de deux sites

- sur les 3.823.589 netpaths de la collection, 228.252 netpaths sont co-cités (soit 5,96% des netpaths). Dans cette collection, presque 6% des pages sont des points d'entrée co-cités qui peuvent être qualifiés par notre méthode.
- sur les 43.462 sites web de la collection, 40.239 sites possèdent des points d'entrée pouvant être classés par la méthode des co-citations, c'est-à-dire 92% des sites. Ce chiffre nous parait très encourageant.

Le calcul de la matrice de co-citation, tel que nous l'avons présenté au chapitre 4 (équation 4.1), n'est pas envisageable avec une méthode universelle de multiplication de matrice. En effet, ce type de méthode alloue un nombre flottant pour chaque élément de la matrice, et permet au mieux de traiter un problème de taille de l'ordre 1.000 avec une centaine de méga-octets de mémoire. Il faut donc se tourner vers une représentation utilisant un seul bit par élément. Ceci est possible dans notre cas puisque notre matrice de citation est binaire. Plusieurs développements successifs réalisés en langage C par des personnes de

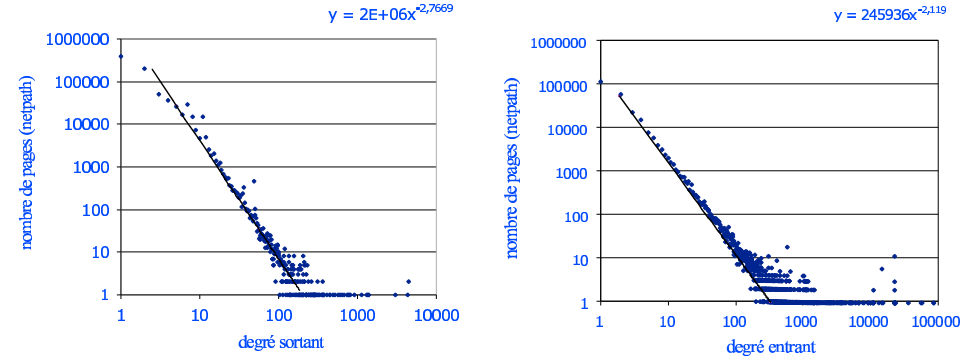

FIG. 6.5 – Distribution des degrés entrant et sortant dans le corpus (pour les netpaths)

notre laboratoire, ont permis d'aboutir à une méthode très rapide de multiplication pour les matrices binaires. Cette méthode de l'ordre de $O(n^3)$ effectue exactement $N_{citant} \times N_{cit}^2$ opérations, où N_{citant} est le nombre de sommets citants pour lesquels $d_s > 1$, et N_{cit} est le nombre de sommets co-cités. Cette méthode testée pour le calcul de la matrice de co-citation de sites (contenant 40.239 sites cités et 12.229 sites citants) donne le résultat de la matrice en une heure avec un ordinateur de puissance moyenne[6]. La matrice de citation de netpaths est 1175 fois plus importante[7], ce qui nous permet une estimation de l'ordre de 1175 heures, soit 48 jours de calcul avec ce même ordinateur. Avec un ordinateur quatre fois plus puissant, le calcul de cette matrice prendrait 12 jours. Le temps de calcul de cette matrice peut paraître excessivement long. Il convient de rappeler que les points d'entrée sont en général accessibles par des URLs stables. Même si le contenu de ces pages est modifié, les propriétés les concernant ne sont généralement pas amenées à évoluer. Ainsi, les métadonnées décrivant ces pages ne nécessitent pas une mise à jour très régulière. Une vérification tous les 6 mois ou même chaque année paraîtrait convenable.

6.4.5 Vers l'extraction de corpus homogènes

Si le calcul de la matrice de co-citation d'une collection importante (de l'ordre de 5 millions de pages) parait envisageable, le découpage par des méthodes de classification automatique ascendante de son graphe induit est une illusion. En effet, ces méthodes de l'ordre de $O(n^3)$ ou de $O(n^2)$ sont coûteuses pour de gros volumes de données. Ainsi, il faut soit s'orienter vers d'autres

[6] 512 méga-octets de mémoire centrale, 1 GHz de vitesse de processeur

[7] Dans la matrice de citation de netpaths $N_{citant} = 446.354$ et $N_{cit} = 228.252$, d'où le rapport entre la taille de la matrice de citation de sites et celle des netpaths : $\frac{446.354 \times (228.252)^2}{12.229 \times (40.239)^2} = 1175$

méthodes de découpage de graphe plus rapides (de l'ordre de $O(n)$), soit envisager l'application de notre approche sur des graphes de co-citation partiels. Une des possibilités consiste, comme pour l'obtention de notre corpus de test, à construire des sous-graphes de co-citation thématiques.

6.4.6 Vers l'intégration de notre approche par des outils de recherche

L'expérience exploratoire sur la collection Wfr4 nous permet de conclure qu'une majorité de sites peuvent être qualifiés par notre approche (environ 92% des sites). Ainsi l'intégration de notre approche par des outils de recherche généralistes est envisageable. Cependant, l'application de nos méthodes d'extraction de corpus homogènes et d'affectation semi-automatique de métadonnées à des corpus de taille importante mériterait une étude approfondie. Enfin, notre approche semi-automatique de caractérisation des points d'entrées sur les sites web fait intervenir un jugement humain. Elle se présente comme un intermédiaire entre les moteurs de recherche dont la volonté est d'indexer l'ensemble du Web par des méthodes automatiques, et les annuaires qui classent et indexent les sites par l'intervention humaine.

Annexe A

Résultats de la qualification manuelle

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
1	http://www.alibaba.online.fr/cybercartes/	entreprise	-	service web	-	-	p. accueil	n.	Cartes virtuelles
2	http://www.alsannuaire.com/	entreprise	-	site de recherche	pages web	annuaire	p. accueil	n.	Annuaire alsacien
3	http://www.alsyd.com/FP/rshift3.html	entreprise	-	homeserveur	-	-	p. contenu	n.	Description d'un logiciel d'astronomie
4	http://www.alyon.asso.fr/generale/histoire/science/	association	-	site de ressources	documents	-	index	n.	Histoire des sciences
5	http://www.anaconda-2.net/andromeda.html	ind.	-	site de ressources	documents	-	index	n.	Dictionnaire encyclopédique d'astronomie
6	http://www.anshare.fr	entreprise	-	site de recherche	logiciels	annuaire	p. accueil	n.	Logithèque
7	http://www.anstj.org/astro	association	-	homeserveur	-	-	p. accueil	a.	Site du télescope Jean marc salomon / Association Nationale Sciences Techniques Jeunesse
8	http://www.arpeges-celestes.com	personne	-	site de ressources	documents	-	p. accueil	n.	Livre d'astronomie (en ligne)
9	http://www.astr.ucl.ac.be/popwork/introclim.html	institution	crt. rech.	homeserveur	-	-	p. contenu	n.	Documents sur la climatologie
10	http://www.astro.ulg.ac.be/~demoulin/glossair.htm	personne	-	homeserveur	-	-	p. contenu	n.	Glossaire d'astronomie
11	http://www.astro.umontreal.ca/groupe	institution	crt. rech.	homeserveur	-	-	p. accueil	a.	Groupe d'astronomie de l'université de Montréal

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
12	http://www.astro.umontreal.ca/~manset/FestivalAstro.html	institution	crt. rech.	homeserveur	-	-	p. contenu	a.	Festival d'astronomie (Canada)
13	http://www.astroclub.ch/	association	club	homeserveur	-	-	p. accueil	a.	Club d'astronomie
14	http://www.astroclub.net/mars/clubjupiter/	association	club	homeserveur	-	-	p. accueil	a.	Club d'astronomie (Canada)
15	http://www.astroclub.net/mercure/centre.astro/	ind.	anim. scient.	homeserveur	-	-	p. accueil	a.	Centre d'astronomie Saint Michel l'observatoire
16	http://www.astroloisir.f2s.com/	personne	-	site de ressources	documents	-	p. accueil	n.	Guide astronomie amateur (Gandoura Mehdi)
17	http://www.astronomag.com/	personne	-	site de ressources	documents	-	p. accueil	n.	Toute l'astronomie en un site par Josselin
18	http://www.astronome.fr/	entreprise	-	homeserveur	-	-	p. accueil	a.	L'astronome : entreprise de matériel d'astronomie
19	http://www.astronomie-paralux.com/	entreprise	-	homeserveur	-	-	p. accueil	a.	Paralux. Entreprise de matériel d'astronomie (optique instrumentale, jumelles, instruments d'astronomie, microscopes)
20	http://www.astronomix.com/	entreprise	-	homeserveur	-	-	p. accueil	a.	Société Astronomix : entreprise de matériel d'astronomie
21	http://www.astrorama.net/	association	anim. scient.	homeserveur	-	-	p. accueil	a.	Astorama : Centre de vulgarisation et d'animation en astronomie

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
22	http://www.astrosurf.com/aca/	association	club	homeserveur	-	-	p. accueil	a.	Astronomie centre Ardennes
23	http://www.astrosurf.com/skylink/doc_astro/pratique/	association	-	site de ressources	documents	-	portail	n.	Skylink : Le site des astronomes amateurs de France (géré par l'association astrotech)
24	http://www.astrosurf.org/astropc/cartes/	personne	-	site de ressources	logiciels	-	p. contenu	n.	Cartes du ciel, programme gratuit d'astronomie - Carte du ciel
25	http://www.astrosurf.org/lcorp/	personne	-	homeserveur	-	-	p. accueil	a.	Site amateur
26	http://www.astrsp-mrs.fr/	institution	crt. rech.	homeserveur	-	-	p. accueil	a.	Observatoire de Provence
27	http://www.atco-fr.com/	association	club	homeserveur	-	-	p. accueil	a.	Association Astronomie Techniques et Communication
28	http://www.atco-fr.com/aec/aec.php3	association	club	homeserveur	-	-	p. accueil	a.	Club d'astronomie en chinonais
29	http://www.auracom.fr/apub/	entreprise	-	homeserveur	-	-	portail	n.	Démonstration logiciel
30	http://www.bde.enseeiht.fr/clubs/astro/	association	club	homeserveur	-	-	p. accueil	a.	Club d'astronomie de l'école N7
31	http://www.bdl.fr/	institution	crt. rech.	homeserveur	-	-	p. accueil	a.	Institut de mécanique céleste
32	http://www.bdl.fr/cnfa/	association	-	homeserveur	-	-	p. accueil	a.	Comité National Français d'Astronomie

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
33	http://www.bdl.fr/webastro.html	institution	crt. rech.	homeserveur	-	-	portail	n.	L'essentiel des ressources d´astronomie sur l'Internet proposé par l'Institut de mécanique céleste
34	http://www.bibli.obspm.fr/catbiba.html	institution	biblioth.	homeserveur	-	-	portail	n.	bibliothèque d'astronomie
35	http://www.bibli.obspm.fr/docastro.html	institution	biblioth.	homeserveur	-	-	portail	n.	Documentation en astronomie-astrophysique et sciences voisines
36	http://www.biblionline.com/Html/annuaire/Astro.html	ind.	-	site de ressources	documents	-	portail	n.	Annuaire d'astronomie et d'astrophysique sur biblio online (e-journal des bibliothèques)
37	http://www.blanchard75.fr/	entreprise	-	homeserveur	-	-	p. accueil	a.	Librairie scientifique
38	http://www.bpi.fr/probib/fracontp/	institution	biblioth.	homeserveur	-	-	portail	n.	Signets d'astronomie de la bibliothèque du Centre Pompidou
39	http://www.burillier-uranie.com/	entreprise	-	homeserveur	-	-	p. accueil	a.	Librairie scientifique (éditeur spécialisé en Astronomie et histoire des sciences)
40	http://www.c-neuf.com/	entreprise	-	site de recherche	pages web	annuaire	p. accueil	n.	Annuaire généraliste gratuit des sites francophones récents
41	http://www.cafe.rapidus.net/algauthi/astro.htm	personne	-	site de ressources	documents	-	index	n.	Portail : rubriques pour comprendre l'astronomie (vulgarisation)

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
42	http://www.cafe.rapidus.net/algauthi/intro.htm	personne	-	site de ressources	documents	-	p. accueil	n.	Site d'astronomie
43	http://www.cala.asso.fr/	association	club	homeserveur	-	-	p. accueil	a.	Club d'astronomie Lyon ampère
44	http://www.cam.org/~sam/	association	société	homeserveur	-	-	p. accueil	a.	Société d'astronomie Montréal
45	http://www.cam.org/~sam/billavf/nineplanets/help.html	association	club	homeserveur	-	-	p. contenu	n.	Glossaire d'astronomie
46	http://www.campus.ecp.fr/astro/	association	club	homeserveur	-	-	p. accueil	a.	Club d'astronomie de l'école Centrale
47	http://www.captifs.org/cera/b9.htm	association	anim. scient.	homeserveur	-	-	p. accueil	a.	Centre d'étude et réalisations astronomique Pegoud (CERAP)
48	http://www.casca.ca/	association	société	homeserveur	-	-	p. accueil	a.	Société canadienne d'astronomie
49	http://www.cc-pays-de-gex.fr/assoc/aorion/	association	club	homeserveur	-	-	p. accueil	a.	Orion : club astronomie du pays de Gex
50	http://www.cc-pays-de-gex.fr/~aorion/	association	club	homeserveur	-	-	p. accueil	a.	Orion : club astronomie du pays de Gex
51	http://www.ccdaude.com/	association	club	homeserveur	-	-	p. accueil	a.	Association des utilisateurs de détecteurs électroniques
52	http://www.centre-congres-toulouse.fr/	institution	-	homeserveur	-	-	p. accueil	a.	Centre des congrès Toulouse

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
53	http://www.chaloop.com/	entreprise	-	site de recherche	pages web	Annuaire	p. contenu	n.	Minissimo : le guide web illusté des meilleurs sites internet
54	http://www.chez.com/astrocam/	personne	-	site de ressources	documents	-	p. accueil	n.	Site amateur d'astronomie
55	http://www.cieletespace.fr/	entreprise	-	homeserveur	-	-	p. accueil	a.	Magazine ciel et espace
56	http://www.circe.fr/pratique/astro/astro_som.html	entreprise	-	site de ressources	documents	-	p. contenu	n.	Carte du ciel, magazine Infoscience
57	http://www.cistemstj.asso.fr	association	anim. scient.	homeserveur	-	-	p. accueil	a.	Cistem
58	http://www.citeweb.net/assem/sect-astro.html	association	-	homeserveur	-	-	p. contenu	a.	Assem (Association Nationale Sciences Techniques Jeunesse) - délégation Provence Alpes Côte d'Azur de l'ANSTJ
59	http://www.clicanoo.com/abcpratique/abcd.asp	entreprise	-	site de recherche	pages web	annuaire	Interface BDD	n.	Clicano : le Portail Ocean-Indien (rubrique informations pratiques)
60	http://www.cnes.fr/	institution	crt. rech.	homeserveur	-	-	p. accueil	a.	Centre national d'études Spatiales (CNES)
61	http://www.cocorico.com/	entreprise	-	site de recherche	pages web	annuaire	p. accueil	n.	Annuaire des produits et services en France (Minitel et Web)
62	http://www.cpod.com/monoweb/asnora/	association	club	homeserveur	-	-	p. accueil	a.	Association normande d'astronomie (ASNORA)
63	http://www.cri.univ-rennes1.fr/scd/adresses.html	institution	biblioth.	homeserveur	-	-	portail	n.	Signets de la biblioth. universitaire de Rennes

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
64	http://www.culture.fr/culture/limoges/astrono.htm	institution	Ministère	homeserveur	-	-	p. contenu	n.	Traité d'astronomie d'Hyginus par Adémar de Chabannes
65	http://www.culturediff.org/	entreprise	-	site de ressources	documents, logiciels	-	p. accueil	a.	Site proposant divers dossiers relatifs à l'histoire des sciences et à l'astronomie égyptienne, ainsi que des logiciels d'astronomie inédits.
66	http://www.cybercable.tm.fr/~tbrahe	ind.	ind.	ind.	ind.	ind.	ind.	ind.	ind.
67	http://www.cyberguide.fr/	entreprise	-	site de recherche	pages web	annuaire	p. accueil	n.	Annuaire de recherche généraliste
68	http://www.cyberus.ca/~ajdesor/astronomie.htm	personne	-	site de ressources	documents	-	portail	n.	Ressources d'astronomie
69	http://www.dasop.obspm.fr/dasop/	institution	crt. rech.	homeserveur	-	-	p. accueil	a.	Département d'astronomie solaire, Observatoire de Paris meudon
70	http://www.dasop.obspm.fr/previ/	institution	crt. rech.	site de ressources	documents	-	p. accueil	n.	Centre de prévision de l'activité solaire de l'observatoire Paris Meudon
71	http://www.dstu.univ-montp2.fr/GRAAL/	institution	crt. rech.	homeserveur	-	-	p. accueil	a.	Groupe de Recherche en Astronomie et Astrophysique du Languedoc
72	http://www.ecila.fr/french/	entreprise	-	site de recherche	pages web	Moteur	p. accueil	n.	Moteur de recherche Ecila

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
73	http://www.ed-phys.fr/books/textes/connaissance2000.html	entreprise	-	homeserveur	-	-	p. contenu	a.	Editeur scientifique : présentation du livre *Eclipse* (commande possible)
74	http://www.edpsciences.com/books/textes/eclipses.html/	entreprise	-	homeserveur	-	-	p. contenu	a.	Editeur scientifique : livre *connaissance des temps* commande possible
75	http://www.education.free.fr/	personne	-	site de recherche	logiciels	annuaire	p. accueil	n.	Le site des logiciels éducatifs diffusés sous licence libre
76	http://www.edunet.ch/classes/c9/espace/	institution	ens. prim.	homeserveur	-	-	index	n.	Projet sur l'espace : *les planètes et les vols habités* - école primaire
77	http://www.ens-lyon.fr/~bgoglin/page_principale.html	personne	-	homeserveur	-	-	p. accueil	a.	page personnelle d'un étudiant de l'ENS
78	http://www.ensta.fr/	institution	ens. sup.	homeserveur	-	-	p. accueil	a.	Ecole Nationale Supérieure de Techniques Avancées (ENSTA)
79	http://www.foorum.fr/	ind.	-	service web	-	-	p. accueil	n.	Dialogues et consultation des newsgroups sur le web
80	http://www.fortunecity.fr/etoiles/voixlactee/6/	ind.	ind.	ind.	ind.	ind.	ind.	ind.	ind.
81	http://www.forum-des-sciences.tm.fr/lieu/planet/planetarium.htm	ind.	anim. scient.	homeserveur	-	-	p. contenu	a.	Planétarium - CCSTI - (centre regional de culture scientifique)

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
82	http://www.fp3.com/ljm/liste_500.asp	personne	-	site de recherche	pages web	annuaire	portail	n.	Liste des glossaires disponibles sur l'Internet
83	http://www.france.diplomatie.fr/label_france/FRANCE/SCIENCES/astro/astro.html	institution	-	site de ressources	documents	-	p. contenu	a.	Journal électronique Label France
84	http://www.franceantiq.fr/slam/latude/latud_fr.htm	entreprise	-	homeserveur	-	-	p. accueil	a.	Librairie livres anciens
85	http://www.freeway.fr/	entreprise	-	site de recherche	pages web	annuaire	p. accueil	n.	Annuaire pays de la Loire
86	http://www.fusl.ac.be/Files/General/BCS/ScTech2.html	institution	biblioth.	homeserveur	-	-	portail	n.	Bibliographie d'orientation
87	http://www.galacticsurf.com/indexF.htm	personne	-	site de recherche	pages web	annuaire	p. accueil	n.	Portail des étoiles
88	http://www.galaxidion.com/	entreprise	-	homeserveur	-	-	p. accueil	a.	Librairie Galaxidion : Le Marché du livre ancien ou épuisé
89	http://www.galaxidion.fr/ecritoire/	entreprise	-	homeserveur	-	-	p. accueil	a.	Librairie
90	http://www.gascogne.fr/Ferme/welcome.htm	association	anim. scient.	homeserveur	-	-	p. accueil	a.	Ferme des étoiles
91	http://www.generation.net/~durand/polaris/	association	club	homeserveur	-	-	p. accueil	a.	Astronomes amateurs Polaris de Lanaudière
92	http://www.geocities.com/CapeCanaveral/Lab/6252/	association	club	homeserveur	-	-	p. accueil	a.	Club d'astronomie Cassiopée de Sillery

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
93	http://www.geocities.com/CapeCanaveral/Lab/9209/astronomie.html	association	club	homeserveur	-	-	portail	n.	Liens vers sites d'astronomie intéressants
94	http://www.geospace.fr/	ind.	-	site de ressources	-	-	p. accueil	n.	Portail grand public sur les Sciences de la terre et de l'Univers
95	http://www.globetrotter.net/astroccd/	association	club	homeserveur	-	-	p. accueil	a.	Premier club d'astronomie 100% virtuel au monde, le Groupe Astro et CCD
96	http://www.globetrotter.net/astronomie_au_quebec/	association	club	site de ressources	documents	-	p. contenu	a.	Consortium, biblioth. et portail de l'astronomie au Québec
97	http://www.graal.univ-montp2.fr/	institution	crt. rech.	homeserveur	-	-	p. accueil	a.	Groupe d'astronomie en Languedoc-Roussillon (GRAL)
98	http://www.guetali.fr/home/thpayet/	personne	-	site de ressources	documents, images	-	p. accueil	n.	Astronomie amateur île de la Réunion
99	http://www.home.ch/~spaw1802/	personne	-	site de ressources	documents	-	p. accueil	n.	Journal d'astronomie - magazine online astronomie et espace
100	http://www.humanite.presse.fr/journal/jour.html	entreprise	-	site de ressources	documents	-	p. accueil	n.	*Humanité* en ligne : Sommaire du jour
101	http://www.iap.fr/saf/lastro.htm	association	-	homeserveur	-	-	p. contenu	n.	Sommaire de la revue *Astronomie*, fondé par Camille Flammarion
102	http://www.iap.fr/sf2a/	association	société	homeserveur	-	-	p. accueil	a.	Société Française d'Astronomie et d'Astrophysique

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
103	http://www.iap.fr/sfsa/	association	société	homeserveur	-	-	p. accueil	a.	Société Française d'Astronomie et d'Astrophysique
104	http://www.ias.fr/web_pronaos/pronaos.html	institution	crt. rech.	homeserveur	-	-	p. accueil	a.	Programme National d'Astronomie Submillimétrique
105	http://www.id-net.fr/~brolis/	personne	-	site de ressources	documents, logiciels	-	p. accueil	n.	*Mon petit Web* documents et logiciels pédagogique spour l'enseignement des sciences physiques et de l'informatique (collège et lycée)
106	http://www.incredibleweather.com/editorial.html	personne	-	site de ressources	documents, images	-	p. accueil	n.	Bew Météo variée
107	http://www.infobiogen.fr/services/deambulum/fr/dictionnaires.html	institution	-	site de ressources	documents	-	portail	n.	Portail : dictionnaires, lexiques scientifiques et techniques
108	http://www.infobourg.com/	entreprise	-	site de ressources	documents	-	p. accueil	n.	Infobourg prof - ejournal (portail) destiné aux professeurs
109	http://www.infoscience.fr/pratique/astro/astro_som.html	entreprise	-	site de ressources	documents	-	p. contenu	n.	Carte du ciel, magazine infoscience
110	http://www.infranet.fr/~captifs/cera/b9.htm	association	anim. scient.	homeserveur	-	-	p. accueil	a.	Cerap - club d'astro de Belfort (observ. Amateur)
111	http://www.inria.fr/ariana/personnel/Andre.Jalobeanu	personne	-	homeserveur	-	-	p. accueil	a.	Page personnelle d'André Jaboneau

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
112	http://www.insu.cnrs-dir.fr/Documentation/Insu_doc/lab_astro.html	institution	crt. rech.	ind.	ind.	ind.	ind.	ind.	ind.
113	http://www.iquebec.com/aaAstronad/	personne	-	homeserveur	-	-	p. accueil	a.	Découvrir l'astronomie, site d'astronomie amateur -
114	http://www.jura.ch/educ/astro/	association	-	homeserveur	-	-	p. accueil	a.	Société jurassienne d'astronomie et observatoire
115	http://www.kyxar.fr/~lingane/index2.html	personne	-	site de ressources	documents	-	p. accueil	n.	Appolo Astronomie
116	http://www.ladepeche.com	entreprise	-	site de ressources	documents	-	p. accueil	n.	La dépêche : site du *Journal du midi*
117	http://www.lexpress.presse.fr/Express/Info/Sciences/	entreprise	-	site de ressources	documents	-	index	n.	site de l'*Express*, index des articles de science
118	http://www.liberation.com/sciences/	entreprise	-	site de ressources	documents	-	index	n.	Index du site *Libération*, rubrique science
119	http://www.linternaute.com/webpassion/webpassionint.shtml	entreprise	-	site de ressources	documents	-	index	n.	Revue de presse du Web
120	http://www.linternaute.fr	entreprise	-	site de recherche	pages web	annuaire	p. accueil	n.	Annuaire *L'internaute, l'Internet tout simplement*
121	http://www.maison-astronomie.com/	entreprise	-	homeserveur	-	-	p. accueil	a.	Maison de l'astronomie : vente de matériel d'astronomie

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
122	http://www.malexism.com/copernic/	personne	-	site de ressources	documents	-	p. accueil	n.	*Copernic* : Titre d'initiation à l'astronomie par Marc-Alexis Morelle (Astronomie, Histoire et Astronautique)
123	http://www.malexism.com/copernic/hist/histoire.html	personne	-	site de ressources	documents	-	index	n.	Site sur l'histoire de l'astronomie
124	http://www.masterouaib.claranet.fr/alastrosom.htm	association	club	homeserveur	-	-	p. accueil	a.	Association lunairienne d'astronomie
125	http://www.math.unicaen.fr/~reyssat/laplace/	personne	-	homeserveur	-	-	p. contenu	n.	Le scientifique Laplace présenté par Eric Reyssat (page personnelle)
126	http://www.maths.univ-rennes1.fr/~rouxph/astro.html	personne	-	homeserveur	-	-	index	n.	Signets d'astronomie par Philippe Roux (page personnelle)
127	http://www.megagiciel.com/243.html	entreprise	-	site de recherche	logiciels	annuaire	index	n.	portail logiciels de sciences
128	http://www.meteo.org/ad-astro.htm	personne	-	site de ressources	documents	-	index	n.	Index de la rubrique astronomie sur le site de Météo de Eve Christian
129	http://www.meteo.org/astridx.htm	personne	-	site de ressources	documents	-	portail	n.	Portail d'astronomie sur le site de météo de Eve Christian
130	http://www.microid.com/maison.htm	entreprise	-	homeserveur	-	-	p. accueil	a.	La maison de l'astronomie :vente de matériel

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
131	http://www.multimania.com/ascala/	association	club	homeserveur	-	-	p. accueil	a.	Association Calédonienne d'Astronomie
132	http://www.multimania.com/cadrael/	association	club	homeserveur	-	-	p. accueil	a.	Cercle Astronomique pour le Développement des Rencontres entre Amateurs.
133	http://www.multimania.com/cesarigd/astro1.htm	personne	-	homeserveur	-	-	p. contenu	a.	Page personnelle astronomie avec de nombreuses photos
134	http://www.netrover.com/~tremblay/	personne	-	homeserveur	-	-	p. accueil	a.	Site amateur d'ovni
135	http://www.nomade.fr/cat/science_technolog/physique/astrophysique_astro/	entreprise	-	site de recherche	pages web	annuaire	index	n.	Annuaire nomade catégorie astronomie
136	http://www.nrpyrenees.com/	entreprise	-	site de ressources	documents	-	p. accueil	n.	*La nouvelle république des Pyrénées*
137	http://www.obs-besancon.fr/	institution	crt. rech.	homeserveur	-	-	p. accueil	a.	Observatoire de Besançon
138	http://www.obs-mip.fr/omp/	institution	crt. rech.	homeserveur	-	-	p. accueil	a.	Observatoire Midi-Pyrénées
139	http://www.obs-nancay.fr/	institution	crt. rech.	homeserveur	-	-	p. accueil	a.	Station de Radioastronomie de Nançay
140	http://www.obs-nice.fr/	institution	crt. rech.	homeserveur	-	-	portail	a.	Observatoire de la côte d'azur
141	http://www.obspm.fr/france.html	institution	crt. rech.	homeserveur	-	-	portail	n.	Portails des sites d'astronomie (site observatoire de Paris)
142	http://www.oceanet.fr/Associations/san/	association	société	homeserveur	-	-	p. accueil	a.	Société d'astronomie de Nantes

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
143	http://www.oceanet.fr/associations/san/	association	société	homeserveur	-	-	p. accueil	a.	Société d'astronomie de Nantes
144	http://www.oma.be/BIRA-IASB/SRBA/	association	société	homeserveur	-	-	p. accueil	a.	Société Royale Belge d'Astronomie, de Météorologie et de Physique du Globe
145	http://www.osco.nb.ca/nbanb/	personne	-	homeserveur	-	-	p. accueil	a.	Page amateur
146	http://www.ot-st-jean-de-monts.fr/musees/biblio.htm	institution	office tourisme	homeserveur	-	-	p. contenu	a.	Site de la biblioth. municipale de Saint Jean de Monts
147	http://www.perigord.tm.fr/astronom/index20.htm	association	club	homeserveur	-	-	p. accueil	a.	L'astronomie amateur en Dordogne, avec la Section Astronomie du Foyer Laïque d'Education Populaire.
148	http://www.peterlang.com	entreprise	-	homeserveur	-	-	p. accueil	a.	Groupe éditorial
149	http://www.philoscience.com/	entreprise	-	homeserveur	-	-	p. accueil	a.	Librairie spécialisée dans les livres anciens et d'occasion
150	http://www.physique.univ-montp2.fr/trois_siecles	ind.	Erreur404	ind.	Erreur404	Erreur404	ind.	ind.	Erreur404
151	http://www.ping.be/eurospace/astro.htm	association	club	homeserveur	-	-	p. accueil	a.	Groupe astronomie de Belgique (GAS)
152	http://www.ping.be/eurospace/stagegas.htm	association	club	homeserveur	-	-	p. contenu	a.	Présentation des stages du GAS (Groupe d'Astronomie de Belgique)
153	http://www.planete-mars.com	association	-	homeserveur	-	-	p. accueil	a.	Projet planète mars

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
154	http://www.planetscapes.com/solar/french/homepage.htm	personne	-	site de ressources	documents	-	p. accueil	n.	*Regards sur le système solaire livre* par Calvin J. Hamilton (livre en ligne)
155	http://www.pleiades.ch	association	club	homeserveur	-	-	p. accueil	a.	Les pléiades - Société d'astronomie de Saint Imier
156	http://www.quebectel.com/astroccd/	association	club	homeserveur	-	-	p. accueil	a.	Groupe Astro et CCD
157	http://www.quellemeteo.com/	ind.	-	site de recherche	pages web	-	p. accueil	n.	Portail Météo
158	http://www.quid.fr/	entreprise	-	site de ressources	documents	-	p. accueil	n.	Site du Quid
159	http://www.radio-astronomie.com/	personne	-	homeserveur	-	-	p. accueil	a.	Site dédié à la radioastronomie d'amateur.
160	http://www.radio-astronomie.com/default.htm	personne	-	homeserveur	-	-	p. accueil	a.	Site dédié à la radioastronomie d'amateur.
161	http://www.radio-astronomie.com/francais.htm	personne	-	homeserveur	-	-	p. accueil	a.	Ce site dédié à la radioastronomie d'amateur.
162	http://www.rasc.ca/srac.html	association	société	homeserveur	-	-	p. accueil	a.	Société Royale d'Astronomie du Canada
163	http://www.recapsite.com/	ind.	-	site de recherche	pages web	annuaire	p. accueil	n.	Annuaire francophone
164	http://www.san-fr.com/	association	société	homeserveur	-	-	p. accueil	a.	Société d'astronomie de Nantes

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
165	http://www.science-tech.nmstc.ca/francais/schoolzone/basesurastronomie.cfm/	institution	Musée	homeserveur	-	-	p. contenu	n.	Notions de base en astronomie
166	http://www.sciences-en-ligne.com/Frames_Dictionary.asp	entreprise	-	site de ressources	documents	-	p. contenu	n.	Dictionnaire scientifique en ligne
167	http://www.selectilinks.com/	entreprise	-	site de recherche	pages web	annuaire	p. accueil	n.	Select link, annuaire francophone du Web gratuit
168	http://www.septembremedia.com/	entreprise	-	homeserveur	-	-	p. contenu	a.	Septembre media, l'école branchée
169	http://www.site-wap.com/	ind.	-	site de ressources	son	-	p. accueil	n.	Wap
170	http://www.solarviews.com/french/homepage.htm	personne	-	site de ressources	documents	-	p. accueil	n.	*Solarview*, livre sur le système solaire
171	http://www.splatsearch.com/lan/index_french.html	entreprise	-	site de recherche	pages web	annuaire	p. accueil	n.	Annuaire : Splat Search
172	http://www.sup.adc.education.fr/bib/rens/FnCadi.htm	institution	gouv	homeserveur	-	-	p. contenu	a.	Centres d'Acquisition et de Diffusion de l'Information Scientifique et Technique
173	http://www.teknea.com/	entreprise	-	homeserveur	-	-	p. accueil	a.	Editeur de livres scientifiques et de produits multimédias (cd-rom)
174	http://www.total.net/~clairmar/cadi/	association	club	homeserveur	-	-	p. accueil	a.	Club astronomie de Drummondville (Canada)

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
175	http://www.total.net/~clairmar/cadi.htm	association	club	homeserveur	-	-	p. accueil	a.	Club astronomie Drummondville
176	http://www.trassud.com	entreprise	-	homeserveur	-	-	p. accueil	a.	Entreprise artisanale de mécanique générale et de chaudronnerie spécialisée dans la fabrication de montures et de matériel astronomique
177	http://www.trifide.com/quasar/	association	club	homeserveur	-	-	p. accueil	a.	groupe d'astronomie Quasar (équipe de 8 passionnés d'astronomie)
178	http://www.trifide.com/quasar/index-2.html	association	club	homeserveur	-	-	index	a.	Groupe d'astronomie Quasar
179	http://www.unige.ch/science-cite/astroqr/qr.html	ind.	ind.	ind.	ind.	ind.	ind.	ind.	ind.
180	http://www.unil.ch/sc/pages/bazar/articles/phys/astronomie/cielnoir.htm	institution	ens. sup.	homeserveur	-	-	p. contenu	n.	Article vulgarisateur *pourquoi la nuit est-elle noire ?* (site de l'université de Lausanne)
181	http://www.univ-lemans.fr/~caum/	association	club	homeserveur	-	-	p. accueil	a.	Club d'astronomie de l'université du Maine
182	http://www.univ-lemans.fr/~caum/eclipse99.html	association	club	homeserveur	-	-	p. contenu	n.	Eclipse totale de soleil
183	http://www.val-de-france.com/	entreprise	-	site de recherche	pages web	annuaire	p. accueil	n.	Portail val de France

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
184	http://www.vianice.fr/	entreprise	-	site de recherche	pages web	annuaire	p. accueil	n.	Portail local de Nice
185	http://www.votations.com/fr/	entreprise	-	service web	-	-	p. accueil	ind.	Insertion de votes, sondages sur les sites
186	http://www.voyager3.com/	association	club	homeserveur	-	-	p. accueil	a.	Club d'astronomie
187	http://www.waponthenet.com/	entreprise	-	site de recherche	Services	annuaire	p. accueil	n.	Annuaire des ressources wap
188	http://www.whoyou.com/	entreprise	-	site de recherche	pages web	annuaire	p. accueil	n.	Annuaire *Whoyou*
189	http://www.yahoo.fr	entreprise	-	site de recherche	pages web	annuaire	p. accueil	n.	Annuaire *Yahoo*
190	http://www2.ac-nice.fr/clea/	institution	-	homeserveur	-	-	p. accueil	a.	Comité de liaison enseignants et astronomes
191	http://www2.cnam.fr/evariste/evariste/8_ADRESS/adresses.htm	institution	anim. scient.	homeserveur	-	-	portail	n.	Quelques bonnes adresses pour découvrir le Web (liens pédagogiques)
192	http://www2.ec-lille.fr/astro/	association	club	homeserveur	-	-	p. accueil	a.	Club d'astronomie de l'école Centrale de Lille
193	http://www3.sympatico.ca/marief.lefebvre/	association	société	homeserveur	-	-	p. accueil	a.	Société d'astronomie du planétarium de Montréal (Canada)
194	http://wwwcenbg.in2p3.fr/Astroparticule/celeste/	institution	crt. rech.	homeserveur	-	-	p. contenu	a.	Projet Céleste, la construction d'un détecteur au sol de rayon gamma
195	http://wwwfirback.ias.u-psud.fr/users/dole/aphelie/	association	club	homeserveur	-	-	p. accueil	a.	Association APHELIE

Num	URL	T-autorité	T-autorité2	T-site	T-site2	T-site3	T-page	T-inf.	Commentaire
196	http://wwwhip.obspm.fr/~arenou/	personne	-	homeserveur	-	-	p. accueil	a.	Page personnelle d'un chercheur astronome
197	http://wwwusr.obspm.fr/commissions/cnap/CNAP.html	institution	-	homeserveur	-	-	p. accueil	a.	Le Conseil National des Astronomes et Physiciens (CNAP)
198	http://yann.duchemin.free.fr/astro/astro.htm	personne	-	homeserveur	-	-	p. accueil	n.	Trucs et astuces pour améliorer l'utilisation de votre matériel d'astronomie

Annexe B

Listes des agrégats obtenus par la méthode du lien complet

(au seuil de coupure le plus haut dans le dendogramme)

Agrégat	Num	T-autorité	T-autorité2	T-site	T-inf.	T-page	Commentaire
Ag1	1	entreprise	-	service web	n.	p. accueil	Cartes virtuelles
Ag1	2	entreprise	-	site de recherche	n.	p. accueil	Annuaire alsacien
Ag10	27	association	club	homeserveur	a.	p. accueil	Association Astronomie Techniques et Communication
Ag10	43	association	club	homeserveur	a.	p. accueil	Club d'astronomie Lyon ampère
Ag10	181	association	club	homeserveur	a.	p. accueil	Club d'astronomie de l'université du Maine
Ag10	30	association	club	homeserveur	a.	p. accueil	Club d'astronomie de école N7
Ag10	124	association	club	homeserveur	a.	p. accueil	Association lunairienne d'astronomie
Ag10	186	association	club	homeserveur	a.	p. accueil	Club astronomie
Ag10	13	association	club	homeserveur	a.	p. accueil	Club astronomie
Ag11	15	Ind.	anim. scient.	homeserveur	a.	p. accueil	Centre d'astronomie Saint Michel l'observatoire
Ag11	21	association	anim. scient.	homeserveur	a.	p. accueil	Astorama : Centre de vulgarisation et d'animation en astronomie
Ag11	81	Ind.	anim. scient.	homeserveur	a.	p. contenu	Planétarium - CCSTI - (centre régional de culture scientifique)
Ag12	17	personne	-	site de ressources	n.	p. accueil	Toute l'astronomie en un site par Josselin Auguste
Ag12	16	personne	-	site de ressources	n.	p. accueil	Guide astronomie amateur (Gandoura Mehdi)
Ag12	122	personne	-	site de ressources	n.	p. accueil	Copernic : Titre d'initiation à l'astronomie par Marc-Alexis Morelle (Astronomie, Histoire et Astronautique)
Ag12	141	institution	crt. rech.	homeserveur	n.	Portail	Portails des sites d'astronomie (site Observatoire de Paris)
Ag13	51	association	club	homeserveur	a.	p. accueil	Association des utilisateurs de détecteurs électroniques
Ag13	194	institution	crt. rech.	homeserveur	a.	p. contenu	Projet Céleste : construction d'un détecteur au sol de rayon gamma

Agrégat	Num	T-autorité	T-autorité2	T-site	T-inf.	T-page	Commentaire
Ag13	121	entreprise	-	homeserveur	a.	p. accueil	Maison de l'astronomie : Matériel
Ag13	18	entreprise	-	homeserveur	a.	p. accueil	L'astronome : entreprise de matériel d'astronomie, grandes marques de matériel d'astronomie
Ag13	19	entreprise	-	homeserveur	a.	p. accueil	Paralux. entreprise de matériel d'astronomie. Spécialiste dans le domaine de l'optique instrumentale, fabrication et commercialisation de jumelles, d'instruments d'astronomie et de microscopes
Ag13	20	entreprise	-	homeserveur	a.	p. accueil	Société Astronomix : entreprise de matériel d'astronomie
Ag14	178	association	club	homeserveur	a.	index	Groupe d'astronomie Quasar - plan du site
Ag14	22	association	club	homeserveur	a.	p. accueil	Astronomie, Centre des Ardennes
Ag15	132	association	club	homeserveur	a.	p. accueil	Cercle Astronomique pour le Développement des Rencontres entre Amateurs.
Ag15	23	association	-	site de ressources	n.	Portail	Skylink : le site des astronomes amateurs de France (géré par l'association Astrotech)
Ag16	25	personne	-	homeserveur	a.	p. accueil	Site amateur
Ag16	133	personne	-	homeserveur	a.	p. contenu	Page personnelle astronomie avec de nombreuses photos
Ag16	196	personne	-	homeserveur	a.	p. accueil	Page personnelle d'un chercheur astronome
Ag16	113	personne	-	homeserveur	a.	p. accueil	Découvrir l'astronomie- Astronomie amateur -
Ag16	77	personne	-	homeserveur	a.	p. accueil	Page personnelle d'un étudiant de l'ENS
Ag17	58	association	-	homeserveur	a.	p. contenu	Assem (Association Nationale Sciences Techniques Jeunesse) - délégation Provence Alpes Côte d'Azur de l'ANSTJ
Ag17	26	institution	crt. rech.	homeserveur	a.	p. accueil	Observatoire de Provence
Ag18	29	entreprise	-	homeserveur	n.	Portail	Démonstration logiciel

Agrégat	Num	T-autorité	T-autorité2	T-site	T-inf.	T-page	Commentaire
Ag18	107	institution	-	site de ressources	n.	Portail	Portail de la biologie et de la bioinformatique
Ag19	139	institution	crt. rech.	homeserveur	a.	p. accueil	Station de Radioastronomie de NANÇAY
Ag19	31	institution	crt. rech.	homeserveur	a.	p. accueil	Institut de mécanique céleste
Ag19	138	institution	crt. rech.	homeserveur	a.	p. accueil	Observatoire Midi-Pyrénées
Ag19	140	institution	crt. rech.	homeserveur	a.	Portail	Observatoire de la côte d'azur
Ag19	60	institution	crt. rech.	homeserveur	a.	p. accueil	CNES
Ag19	69	institution	crt. rech.	homeserveur	a.	p. accueil	Département d'astronomie solaire - Observatoire de Paris Meudon
Ag19	70	institution	crt. rech.	site de ressources	n.	p. accueil	Centre de prévision de l'activité solaire de l'Observatoire Paris Meudon
Ag19	137	institution	crt. rech.	homeserveur	a.	p. accueil	Observatoire de Besançon (index)
Ag2	3	entreprise	-	homeserveur	n.	p. contenu	Description logiciel
Ag2	24	personne	-	site de ressources	n.	p. contenu	Cartes du ciel, programme gratuit d'astronomie - Carte du ciel
Ag20	197	institution	-	homeserveur	a.	p. accueil	Le Conseil National des Astronomes et Physiciens (CNAP)
Ag20	32	association	-	homeserveur	a.	p. accueil	Comité National Français d'Astronomie
Ag21	35	institution	Bibliothèque	homeserveur	n.	Portail	Documentation en astronomie-astrophysique et sciences voisines
Ag21	34	institution	Bibliothèque	homeserveur	n.	Portail	Bibliothèques d'astronomie
Ag22	89	entreprise	-	homeserveur	a.	p. accueil	Librairie
Ag22	88	entreprise	-	homeserveur	a.	p. accueil	Galaxidion : le Marché du livre ancien ou épuisé
Ag22	84	entreprise	-	homeserveur	a.	p. accueil	Librairie livres anciens
Ag22	37	entreprise	-	homeserveur	a.	p. accueil	Librairie scientifique
Ag23	192	association	club	homeserveur	a.	p. accueil	Club d'astronomie de l'école Centrale de Lille
Ag23	39	entreprise	-	homeserveur	a.	p. accueil	Librairie scientifique

Agrégat	Num	T-autorité	T-autorité2	T-site	T-inf.	T-page	Commentaire
Ag23	159	personne	-	homeserveur	a.	p. accueil	Ce site est entièrement dédié à la radioastronomie amateur.
Ag23	131	association	club	homeserveur	a.	p. accueil	Association calédonienne d'astronomie
Ag23	176	entreprise	-	homeserveur	a.	p. accueil	Entreprise artisanale de mécanique générale et de chaudronnerie spécialisée dans la fabrication de montures et de matériel astronomique
Ag23	160	personne	-	homeserveur	a.	p. accueil	Site entièrement dédié à la radioastronomie amateur.
Ag24	67	entreprise	-	site de recherche	n.	p. accueil	*Cyberguide* Annuaire de recherche généraliste
Ag24	40	entreprise	-	site de recherche	n.	p. accueil	*C-neuf* annuaire généraliste gratuit des sites francophones récents
Ag24	53	entreprise	-	site de recherche	n.	p. contenu	*Minissimo*, le guide web illusté des meilleurs sites internet
Ag24	188	entreprise	-	site de recherche	n.	p. accueil	*Whoyou*, annuaire de sites Web
Ag24	167	entreprise	-	site de recherche	n.	p. accueil	*Select link* - annuaire francophone du Web gratuit
Ag25	41	personne	-	site de ressources	n.	index	Index vers des rubriques pour comprendre l'astronomie (vulgarisation)
Ag25	123	personne	-	site de ressources	n.	index	Site sur l'histoire de l'astronomie
Ag25	170	personne	-	site de ressources	n.	p. accueil	*Solar view*, livre sur le système solaire
Ag25	154	personne	-	site de ressources	n.	p. accueil	*Regards sur le système solaire* livre par Calvin J. Hamilton (livre en ligne)
Ag26	126	personne	-	homeserveur	n.	index	Signets d'astronomie par Philippe Roux (page personnelle)
Ag26	54	personne	-	site de ressources	n.	p. accueil	Site amateur d'astronomie Site
Ag26	42	personne	-	site de ressources	n.	p. accueil	Site astronomie
Ag26	161	personne	-	homeserveur	a.	p. accueil	site dédié à la radioastronomie d'amateur.
Ag27	92	association	club	homeserveur	a.	p. accueil	Club d'astronomie Cassiopée de Sillery
Ag27	175	association	club	homeserveur	a.	p. accueil	Club d'astronomie Drummondville

Agrégat	Num	T-autorité	T-autorité2	T-site	T-inf.	T-page	Commentaire
Ag27	91	association	club	homeserveur	a.	p. accueil	Astronomes amateurs Polaris de Lanaudière
Ag27	44	association	société	homeserveur	a.	p. accueil	Société d'astronomie de Montréal
Ag28	155	association	club	homeserveur	a.	p. accueil	Les pléiades - Société d'astronomie de Saint Imier
Ag28	46	association	club	homeserveur	a.	p. accueil	Club d'astronomie école Centrale
Ag28	102	association	société	homeserveur	a.	p. accueil	Société Française d'Astronomie et d'Astrophysique
Ag28	112	institution	crt. rech.	Ind.	Ind.	Ind.	Ind.
Ag28	164	association	société	homeserveur	a.	p. accueil	Société d'astronomie de Nantes
Ag28	143	association	société	homeserveur	a.	p. accueil	Société d'astronomie de Nantes
Ag28	103	association	société	homeserveur	a.	p. accueil	Société Française d'Astronomie et d'Astrophysique
Ag28	142	association	société	homeserveur	a.	p. accueil	Société d'astronomie de Nantes
Ag29	110	association	anim. scient.	homeserveur	a.	p. accueil	Cerap - club d'astronomie de Belfort (observatoire Amateur)
Ag29	47	association	anim. scient.	homeserveur	a.	p. accueil	Centre d'étude et réalisations astronomiques Pegoud- CERAP
Ag3	57	association	anim. scient.	homeserveur	a.	p. accueil	Cistem
Ag3	4	association	-	site de ressources	n.	index	Histoire des sciences
Ag30	68	personne	-	site de ressources	n.	Portail	Ressources astronomie
Ag30	162	association	société	homeserveur	a.	p. accueil	Société Royale d'Astronomie du Canada
Ag30	48	association	société	homeserveur	a.	p. accueil	Société canadienne d'astronomie
Ag31	49	association	club	homeserveur	a.	p. accueil	Orion Club d'astronomie du pays de Gex
Ag31	50	association	club	homeserveur	a.	p. accueil	Orion Club d'astronomie du pays de Gex
Ag32	56	entreprise	-	site de ressources	n.	p. contenu	Carte du ciel, magazine *Infoscience*
Ag32	109	entreprise	-	site de ressources	n.	p. contenu	Carte du ciel, magazine *Infoscience*
Ag33	189	entreprise	-	site de recherche	n.	p. accueil	Annuaire *Yahoo*

Agrégat	Num	T-autorité	T-autorité2	T-site	T-inf.	T-page	Commentaire
Ag33	72	entreprise	-	site de recherche	n.	p. accueil	Moteur de recherche *Ecila*
Ag33	61	entreprise	-	site de recherche	n.	p. accueil	Annuaire des produits et service en France (Minitel et Web)
Ag34	62	association	club	homeserveur	a.	p. accueil	Association Normande d'Astronomie
Ag34	198	personne	-	homeserveur	n.	p. accueil	Trucs et astuces pour améliorer l'utilisation de votre matériel d'astronomie
Ag35	64	institution	Ministère	homeserveur	n.	p. contenu	Traité d'astronomie d'Hyginus
Ag35	86	institution	Bibliothèque	homeserveur	n.	Portail	Bibliographie d'orientation
Ag36	66	Ind.	Ind.	Ind.	Ind.	Ind.	Ind.
Ag36	147	association	club	homeserveur	a.	p. accueil	L'astronomie amateur en Dordogne, avec la Section Astronomie du Foyer Laïque d'Education Populaire.
Ag37	73	entreprise	-	homeserveur	a.	p. contenu	Editeur scientifique : livre *Eclipse* présentation et commande
Ag37	74	entreprise	-	homeserveur	a.	p. contenu	Editeur scientifique : livre *Connaissance des temps* présentation et commande
Ag38	180	institution	ens. sup.	homeserveur	n.	p. contenu	Article vulgarisateur *Pourquoi la nuit est-elle noire ?* sur le site le l'université de Lausanne
Ag38	76	institution	ens. prim.	homeserveur	n.	index	Projet sur l'Espace, les Planètes et les Vols habités - école primaire
Ag39	80	Ind.	Ind.	Ind.	Ind.	Ind.	Ind.
Ag39	99	personne	-	site de ressources	n.	p. accueil	Journal d'astronomie - magazine online astronomie et espace
Ag4	71	institution	crt. rech.	homeserveur	a.	p. accueil	Groupe de Recherche en Astronomie et Astrophysique du Languedoc
Ag4	5	Ind.	-	site de ressources	n.	index	Dictionnaire encyclopédique d'astronomie
Ag4	97	institution	crt. rech.	homeserveur	a.	p. accueil	Groupe d'astronomie en Languedoc - Roussillon (GRAL)

Agrégat	Num	T-autorité	T-autorité2	T-site	T-inf.	T-page	Commentaire
Ag4	55	entreprise	-	homeserveur	a.	p. accueil	Magazine *Ciel et espace*
Ag40	85	entreprise	-	site de recherche	n.	p. accueil	Annuaire pays de la Loire
Ag40	183	entreprise	-	site de recherche	n.	p. accueil	Portail val de France
Ag41	87	personne	-	site de recherche	n.	p. accueil	Portail des étoiles
Ag41	118	entreprise	-	site de ressources	n.	index	Index du site libération rubrique science
Ag42	156	association	club	homeserveur	a.	p. accueil	Groupe Astro et CCD
Ag42	95	association	club	homeserveur	a.	p. accueil	Premier club d'astronomie 100% virtuel au monde, le Groupe Astro et CCD f
Ag42	130	entreprise	-	homeserveur	a.	p. accueil	La maison de l'astronomie : vente de matériel
Ag43	96	association	club	site de ressources	a.	p. contenu	Consortium, bibliothèque et portail d'astronomie au Québec
Ag43	129	personne	-	site de ressources	n.	Portail	Portail d'astronomie sur le site de météo de Eve Christian
Ag44	157	Ind.	-	site de recherche	n.	p. accueil	Portail Météo
Ag44	135	entreprise	-	site de recherche	n.	index	Annuaire *Nomade* catégorie astronomie
Ag44	100	entreprise	-	site de ressources	n.	p. accueil	*Humanité* en ligne : Sommaire du jour
Ag45	191	institution	anim. scient.	homeserveur	n.	Portail	Quelques bonnes adresses pour découvrir le Web (pédagogiques)
Ag45	105	personne	-	site de ressources	n.	p. accueil	*Mon petit Web* documents et logiciels pédagogiques pour l'enseignement des sciences physiques et de l'informatique (collège et lycée)
Ag46	168	entreprise	-	homeserveur	a.	p. contenu	Septembre media - école branchée
Ag46	108	entreprise	-	site de ressources	n.	p. accueil	Infobourg prof - ejournal (portail) destine aux profs (édité par Septembre média)
Ag47	115	personne	-	site de ressources	n.	p. accueil	Appolo Astronomie
Ag47	125	personne	-	homeserveur	n.	p. contenu	Le scientifique Laplace présenté par Eric Reyssat (page personnelle)
Ag48	136	entreprise	-	site de ressources	n.	p. accueil	*La nouvelle république des Pyrénées*

Agrégat	Num	T-autorité	T-autorité2	T-site	T-inf.	T-page	Commentaire
Ag48	116	entreprise	-	site de ressources	n.	p. accueil	*La dépêche*, site du journal du midi
Ag49	120	entreprise	-	site de recherche	n.	p. accueil	L'internaute - internet tout simplement (logiciels, documents)
Ag49	119	entreprise	-	site de ressources	n.	index	Revue de presse du Web
Ag5	75	personne	-	site de recherche	n.	p. accueil	Le site des logiciels éducatifs diffusés sous licence libre
Ag5	6	entreprise	-	site de recherche	n.	p. accueil	Logithèque
Ag50	151	association	club	homeserveur	a.	p. accueil	Groupe astronomie de Belgique (GAS)
Ag50	144	association	société	homeserveur	a.	p. accueil	Société Royale Belge d'Astronomie, de Météorologie et de Physique du Globe
Ag50	177	association	club	homeserveur	a.	p. accueil	groupe d'astronomie Quasar (équipe de 8 passionnés d'astronomie)
Ag51	148	entreprise	-	homeserveur	a.	p. accueil	Groupe éditorial
Ag51	173	entreprise	-	homeserveur	a.	p. accueil	Editeur de livres scientifiques et de produits multimédias (cd-rom)
Ag52	158	entreprise	-	site de ressources	n.	p. accueil	Site du Quid
Ag52	166	entreprise	-	site de ressources	n.	p. contenu	Dictionnaire scientifique en ligne
Ag53	169	Ind.	-	site de ressources	n.	p. accueil	Wap
Ag53	187	entreprise	-	site de recherche	n.	p. accueil	Annuaire ressources wap
Ag54	190	institution	-	homeserveur	a.	p. accueil	Comité de liaison enseignants et astronomes
Ag54	179	Ind.	Ind.	Ind.	Ind.	Ind.	Ind.
Ag6	65	entreprise	-	site de ressources	a.	p. accueil	Site proposant divers dossiers relatifs à l'histoire des sciences et à l'astronomie égyptienne, ainsi que des logiciels d'astronomie inédits.
Ag6	8	personne	-	site de ressources	n.	p. accueil	Livre d'astronomie
Ag7	10	personne	-	homeserveur	n.	p. contenu	Glossaire astro

Agrégat	Num	T-autorité	T-autorité2	T-site	T-inf.	T-page	Commentaire
Ag7	45	association	club	homeserveur	n.	p. contenu	Glossaire astro
Ag8	36	Ind.	-	site de ressources	n.	Portail	Annuaire astronomie astrophysique sur *biblio online* (ejournal des bibliothèques)
Ag8	11	institution	crt. rech.	homeserveur	a.	p. accueil	Groupe d'astronomie de l'université Montréal
Ag9	134	personne	-	homeserveur	a.	p. accueil	Site amateur d'ovni
Ag9	14	association	club	homeserveur	a.	p. accueil	Club d'astronomie (Canada)
Ag9	174	association	club	homeserveur	a.	p. accueil	Club d'astronomie Drummondville (Canada)
Ag9	193	association	société	homeserveur	a.	p. accueil	société astronomie du planétarium de Montréal (Canada)
Ag9	12	institution	crt. rech.	homeserveur	a.	p. contenu	Festival Astronomie (Canada)

Bibliographie

[dub, 2003] (2003). Dublin core metadata initiative (dcmi). *at http ://dublincore.org, consulté en février 2003.*

[rfc, 2004] (2004). Internet requests for comments (rfc) 2396. *consulté le 1/7/4 at http ://www.faqs.org/rfcs/rfc2396.html.*

[Aguiar, 2002] Aguiar, F. (Juin 2002). *Modélisation d'un système de recherche d'information pour les systèmes hypertextes. Application à la recherche d'information sur le World-Wide-Web.* Thèse de doctorat, Ecole Nationale Supérieure des Mines de Saint Etienne, France.

[Albert et al., 1999] Albert, R., Jeong, H., and Barabasi, A.-L. (1999). Diameter of the World Wide Web. *Nature*, 401 :130–131.

[Atlan, 1979] Atlan, H. (1979). *Entre le cristal et la fumée, Essai sur l'organisation du vivant.* Seuil.

[Balpe et al., 1996] Balpe, J., Lelu, A., Papy, F., and Saleh, I. (1996). *Techniques avancées pour l'hypertexte.* éditions Hermès.

[Bar-Ilan, 2001] Bar-Ilan, J. (2001). How much information the search engines disclose on the links to a web page ? a case study of the Cybermetrics home page. In *Proceedings of the 8th International Conference on Scientometrics and Infometrics, ISSI 2001, Sydney, Australia*, pages 63–73.

[Barnes, 1954] Barnes, J. (1954). Class and committees in a norwegian island parish. *Human relations*, 7 :39–58.

[Bassecoulard-Zitt and Zitt, 1998] Bassecoulard-Zitt, E. and Zitt, M. (1998). La scientométrie : les facettes d'un miroir. *Journal interne de l'INRA.*

[Belkin et al., 1982] Belkin, N., Oddy, R., and Brooks, H. (1982). Ask for information retrieval : Part i. background and theroy. *Journal of Documentation*, 38(2) :61–71.

[Bennouas et al., 2003] Bennouas, T., Bouklit, M., and de Montgolfier, F. (2003). Un modèle gravitationnel du web. In *Premières journées francophones de la Toile, JFT03.*

[Benzecri, 1981] Benzecri, J.-P. (1981). *L'analyse de données (tome 1 et 2).* Dunod, Paris.

[Bharat and Broder, 1998] Bharat, K. and Broder, A. (1998). Measuring the Web.

[Bharat et al., 2000] Bharat, K., Broder, A., Dean, J., and Henzinger, M. (2000). A comparison of techniques to find mirrored hosts on the www. *IEEE Data Engineering Bulletin*, 23(4) :21–26.

[Bharat et al., 1998] Bharat, K., Broder, A., Henzinger, M., Kumar, P., and Venkatasubramanian, S. (1998). Connectivity server : Fast access to linkage information on the Web. In *Proceeding of the Seventh International World Wide Web Conference*.

[Boutin, 1999] Boutin, E. (1999). *Le traitement d'une information massive par l'analyse réseau : méthode, outils, applications*. Thèse de doctorat, Université de droit, d'économie et des sciences d'Aix-Marseille.

[Bradford, 1934] Bradford, S. (1934). Sources of information on specific subjects. *Engineering*, 137 :85–86.

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *WWW7 - Computer Networks and Systems 30 (7)*, pages 107–117. IW3C2.

[Broder et al., 2000] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the Web. In *9th International World Wide Web Conference (WWW9)*, pages 309–320, Amsterdam.

[Bush, 1945] Bush, V. (1945). As we may think. *The Atlantic Monthly*, 1(176) :101–108.

[Case and Higgins, 2000] Case, D. and Higgins, G. (2000). How can we investigate citation behavior ? a study of reasons for citing literature in communication. *Journal of the American Society for Information Science*, 51(7) :635–645.

[Chakrabarti et al., 1998a] Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P., and Rajagopalan, S. (1998a). Automatic resource list compilation by analyzing hyperlink structure and associated text. In *Proceedings of the 7th International World Wide Web Conference*.

[Chakrabarti et al., 1998b] Chakrabarti, S., Dom, B., and Indyk, P. (1998b). Enhanced hypertext categorization using hyperlinks. In *ACM SIGMOD*.

[Chekuri et al., 1996] Chekuri, C., Goldwasser, M., Raghavan, P., and Upfal, E. (1996). Web search using automatic classification. In *Proceedings of WWW-96, 6th International Conference on the World Wide Web*, San Jose, US.

[Chu, 2004] Chu, H. (2004). Taxonomy of inlinked web entities : What does it imply for webometric research ? *Library and Information Science Research : An International Journal*, (In press.).

[Cronin, 1981] Cronin, B. (1981). Agreement and divergence on referencing practice. *Journal of Information Science*, 3(1) :27–33.

[Cronin, 2001] Cronin, B. (2001). Bibliometrics and beyond ; some thoughts on web-based citation analysis. *Journal of Information Science*, 27(1) :1–7.

[Crowston and Williams, 2000] Crowston, K. and Williams, M. (2000). Reproduced and emergent genres of communication on the World Wide Web. *The Information Society*, 16(3) :201–215.

[de Solla Price, 1963a] de Solla Price (1963a). *Science et Suprascience*. Traduction française de G. Lévy, Paris, Fayard, 1972, p.83. Version original en anglais : Little Science, Big Science, New York, Columbia University Press, 118p.

[de Solla Price, 1963b] de Solla Price, D. J. (1963b). *Little Science, Big Science*. Columbia, New York.

[de Solla Price, 1965] de Solla Price, D. J. (1965). Network of scientific papers.

[de Solla Price, 1970] de Solla Price, D. J. (1970). Citation measures of hard science, soft science, technology, and non-science. In Nelson, C. and Pollock, D., editors, *Communication among scientists and engineers*, pages 3–22. Health, Lexington (Mass.).

[Ding et al., 2000] Ding, J., Gravano, L., and Shivakumar, N. (2000). Computing geographical scopes of web resources. In *Proceedings of the 26th International Conference on Very Large Databases (VLDB 2000)*.

[Egghe, 2000] Egghe, L. (2000). New informetric aspects of the Internet : some reflections, many problems. *Journal of Information Science*, 26(5) :329–335.

[Egghe and Rousseau, 1990] Egghe, L. and Rousseau, R. (1990). *Introduction to Informetrics*. Elsevier, Amsterdam.

[Flake et al., 2002] Flake, G. W., Lawrence, S., Giles, C. L., and Coetzee, F. (2002). Self-organization of the Web and identification of communities. *IEEE Computer*, 35(3) :66–71.

[Freeman, 1979] Freeman, L. (1979). Centrality in social networks : I. Conceptual clarification. *Social Networks*, 1 :215–239.

[Garfield, 1965] Garfield, E. (1965). Can citation indexing be automated ? In *Statistical association methods for mechanized documentation : Symposium proceedings*, pages 189–192. Washington, DC : National Bureau of standards.

[Garfield, 1972] Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, (178) :471–479.

[Gibson et al., 1998] Gibson, D., Kleinberg, J., and Raghavan, P. (1998). Inferring web communities from link topology. In *Proc. 9th ACM Conference on Hypertext and Hypermedia*.

[Gilbert, 1977] Gilbert, G. (1977). Referencing as persuasion. *Social Studies of Science*, 7 :113–122.

[Glover et al., 2001] Glover, E., Flake, G., Lawrence, S., Birmingham, W. P., Kruger, A., Giles, C. L., and Pennock, D. (2001). Improving category specific web search by learning query modifications. In *Symposium on Applications and the Internet, SAINT*, pages 23–31, San Diego, CA. IEEE Computer Society, Los Alamitos, CA.

[Gravano, 2000] Gravano, L. (2000). Characterizing web resources for improved search. In *Position paper for the First NSF-DELOS Workshop on Information Seeking, Searching, and Querying in Digital Libraries*.

[Hartigan, 1975] Hartigan, J. A. (1975). *Clustering Algorithms*. John Willey, New York.

[Hicks, 1987] Hicks, D. (1987). Limitations of the co-citation analysis / bibliometric modelling : a reply to franklin. *Social Studies of Science*, 17 :295–316.

[Ingwersen, 1998] Ingwersen, P. (1998). The calculation of web impact factors. *Journal of Documentation*, 54(2) :236–243.

[Kessler, 1963] Kessler, M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14 :10–25.

[Kim, 2000] Kim, H. (2000). Motivations for hyperlinking in scholarly articles : a qualitative study. *Journal of the American Society for Information Science*, 51(10) :887–899.

[Kleinberg, 1999] Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5) :604–632.

[Kumar et al., 1999] Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Trawling the Web for emerging cyber-communities. In *Proceedings of the Eighth World Wide Web Conference.*

[Kwasnik et al., 2001] Kwasnik, B., Crowston, K., Nilan, M., and Roussinov, D. (2001). Identifying document genre to improve Web search effectiveness. *The Bulletin of the American Society for Information Science and Technology*, 27(2).

[Lafouge, 1991] Lafouge, T. (1991). Problématique de la circulation de l'information. *Documentaliste*, 28(3) :132–133.

[Larson, 1996] Larson, R. (1996). Bibliometrics of the World Wide Web : An exploratory analysis of the intellectual structure of Cyberspace. In *Proceedings of the Annual Meeting of the American Society of Information Science*, Baltimore.

[Lawrence, 2001] Lawrence, S. (2001). Online or invisible. *Nature*, 411(6837) :521.

[Lawrence et al., 1999] Lawrence, S., Bollacker, K., and Giles, C. L. (1999). Indexing and retrieval of scientific literature. In *Eighth International Conference on Information and Knowledge Management, CIKM 99*, pages 139–146, Kansas City, Missouri.

[Lawrence and Giles, 1998] Lawrence, S. and Giles, C. L. (1998). Searching the World Wide Web. *Science*, 280 :98–100.

[Lawrence and Giles, 1999] Lawrence, S. and Giles, C. L. (1999). Accessibility of information on the Web. *Nature*, 400(6740) :107–109.

[Lotka, 1926] Lotka, A. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16 :317–323.

[Lévy, 1990] Lévy, P. (1990). *Les technologies de l'intelligence.* Editions de la Découverte.

[Marchiori, 1998] Marchiori, M. (1998). The limits of web metadata and beyond. In *Proceedings of the Seventh International WWW Conference.* IW3C2.

[Margalef, 1958] Margalef, R. (1958). Information theory in ecology. *General Systems*, (3) :36–71.

[Marshakova, 1973] Marshakova, I. V. (1973). Document coupling system based on references taken from science citation index. *Russia, Nauchno - Tekhnicheskaya Informatsiya*, 2(6,3).

[Martyn, 1964] Martyn, J. (1964). Bibliographic coupling. *Journal of Documentation*, 20(4) :236.

[Michelet, 1988] Michelet, B. (1988). *L'analyse des associations*. Thèse de doctorat, Université de Paris VII, UFR de Chimie, Paris.

[Moreno, 1934] Moreno, J. L. (1934). *Fondements de la sociométrie (traduction française de : Who shall survive ? : Foundations of sociometry)*. P.U.F Paris (traduction en 1970).

[Olsina et al., 2001] Olsina, L., , Godoy, D., Lafuente, G., and Rossi, G. (2001). Specifying quality characteristics and attributes for websites. *Lecture Notes in Computer Science*, 2016.

[Perenon, 2000] Perenon, P. (2000). *Réalisation d'un prototype de système de recherche d'informations scientifiques : indexation non thématique sous forme de métadonnées et développement d'une interface de consultation prenant en compte les profils des utilisateurs*. Mémoire de DEA, Sous la direction de Sylvie Lainé-Cruzel, Laboratoire RECODOC, Université Claude Bernard Lyon 1.

[Pinski and Narin, 1976] Pinski, G. and Narin, F. (1976). Citation influence for journal aggregates of scientific publications : Theory, with application to the literature of Physics. *Information Processing and Management*, 12 :297–312.

[Pirolli et al., 1996] Pirolli, P., Pitkow, J., and Rao, R. (1996). Silk from a sow's ear : Extracting usable structures from the Web. In *Proc. ACM Conf. Human Factors in Computing Systems, CHI*. ACM Press.

[Pitkow and Pirolli, 1997] Pitkow, J. and Pirolli, P. (1997). Life, death and lawfulness on the electronic frontier. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing System, CHI'97*, pages 118–125.

[Polanco, 1995] Polanco, X. (1995). Aux sources de la scientométrie. In *Les sciences de l'information : bibliométrie, scientométrie, infométrie. Solaris, (2)*. Presses Universitaire de Rennes.

[Prime et al., 2002a] Prime, C., Bassecoulard, E., and Zitt, M. (2002a). Co-citations and co-sitations : a cautionary view on an analogy. *Scientometrics*, 54(2) :291–308.

[Prime et al., 2002b] Prime, C., Beigbeder, M., and Lafouge, T. (2002b). Clusterisation du web en vue d'extraction de corpus homogènes. In *Actes du 20ème congrès INFORSID*, pages 229–242, Nantes.

[Pritchard, 1969] Pritchard, A. (1969). Statistical bibliography or Bibliometrics ? *Journal of Documentation*, 25(4) :348–349.

[R and Blockeel, 2000] R, R. K. and Blockeel, H. (2000). Web mining research : A survey. In *SIGKDD Explorations*, volume 2, pages 1–15.

[Richy, 2002] Richy, H. (2002). Métadonnées et documents numériques. *Techniques de l'ingénieur. Traité informatique*, H7(155) :1–14.

[Rostaing et al., 1999] Rostaing, H., Boutin, E., and Mannina, B. (1999). Evaluation of internet resources : Bibliometric techniques applications. In *Cybermetrics99*, Colima.

[Rousseau, 1997] Rousseau, R. (1997). Sitations : an exploratory study. In *Cybermetrics*, volume 1.

[Sabidussi, 1966] Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31 :581–603.

[Shannon, 1948] Shannon, C. (1948). A mathematical theory of communications. *Bell System technical Journal*, (27) :379–423, 623–656.

[Small, 1973] Small, H. (1973). Co-citation in the scientific literature. *Journal of the American Society for Information Science*, 24 :265–269.

[Small and Sweeney, 1985] Small, H. and Sweeney, E. (1985). Clustering the science citation index using co-citation. I A comparison of methods.

[Teasdale, 1995] Teasdale, G. (1995). L'hypertexte : historique et applications en bibliothéconomie. *Cursus*, 1(1) :101–108.

[Thelwall, 2003] Thelwall, M. (2003). What is this link doing here ? beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, 8(3).

[Turenne, 2000] Turenne, N. (2000). *Apprentissage statistique pour l'extraction de concepts à partir de textes. Application au filtrage d'informations textuelles.* Thèse de doctorat, Université Louis-Pasteur Strasbourg, ENSAIS.

[van Risjbergen, 1979] van Risjbergen, C. (1979). *Information Retrieval*. Butterworths.

[Wasserman and Faust, 1994] Wasserman, S. and Faust, K. (1994). *Social network analysis : methods and applications*. Cambridge university Press.

[White and Griffith, 1981] White, H. and Griffith, B. (1981). Author cocitation : A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3) :163–172.

[Zipf, 1949] Zipf, G. (1949). *Human Behavior and the Principle of Least Effort : an Introduction to Human Ecology*. Reading, Mass : Addison-Wesley.

[Zitt and Bassecoulard, 1996] Zitt, M. and Bassecoulard, E. (1996). Reassessment of co-citation methods for science indicators : effect of methods improving recall rates. *Scientometrics*, 37(2) :223–244.

[Zitt and Bassecoulard, 1998] Zitt, M. and Bassecoulard, E. (1998). Méthodes de structuration pour l'analyse stratégique des univers scientifiques : les techniques de citation. In *Actes VSST'98 (Veille Stratégique, Scientifique et technologique)*, pages 31–41, Toulouse, France.